SEMI-SUPERVISED LEARNING FOR SPEECH RECOGNITION IN THE CONTEXT OF ACCENT ADAPTATION

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

Accented speech that is under-represented in the training data still suffers high Word Error Rate (WER) with state-of-the-art Automatic Speech Recognition (ASR) systems. Careful collection and transcription of training data for different accents can address this issue, but it is both time consuming and expensive. However, for many tasks such as broadcast news or voice search, it is easy to obtain large amounts of audio data from target users with representative accents, albeit without accent labels or even transcriptions. Semi-supervised training have been explored for ASR in the past to leverage such data, but many of these techniques assume homogeneous training and test conditions. In this paper, we experiment with cross-entropy based speaker selection to adapt a source recognizer to a target accent in a semi-supervised manner, using additional data with no accent labels. We compare our technique to self-training based only on confidence scores and show that we obtain significant improvements over the baseline by leveraging additional unlabeled data on two different tasks in Arabic and English.

Index Terms— Semi-supervised learning, Automatic speech recognition, Accent adaptation

1. INTRODUCTION

Recent ASR systems used in commercial applications are trained on massive amounts of training data for languages such as English. However, such large training corpora usually do not have a balanced distribution of different accents present in the target population. Hence, performance of these systems drops significantly for accented speech, that is not well-represented in the training data [1, 2]. One solution is to adapt the source ASR system to build accent-dependent systems for the major accents, with data collected from the customers using the application [3]. Supervised accent adaptation requires annotating the recorded audio data with accent labels in addition to transcribing it, which is infeasible. We explore semi-supervised learning in this paper to select appropriate training data from a diverse dataset for the goal of target accent adaptation.

Semi-supervised learning has become attractive in ASR given the high cost of transcribing audio data. Self-training is a commonly used technique for semi-supervised learning in speech recognition [4, 5, 6, 7, 8], 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 [9, 10]. 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.

We employ 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.

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 [11]. Since we focus only on acoustic variations in this work, we use the term 'Accent' throughout this paper.

2. SEMI-SUPERVISED LEARNING

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?

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 [12]. The selection can be done at utterance, speaker or session level. The average confident score for the appropriate level is calculated as

$$CS_{\mathcal{S}} = \frac{\sum_{\mathcal{W} \epsilon \mathcal{S}} C_{\mathcal{W}} T_{\mathcal{W}}}{\sum_{\mathcal{W} \epsilon \mathcal{S}} T_{\mathcal{W}}} \tag{1}$$

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 [13, 14]. 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.

2.2. Cross-entropy based data selection

In this section, we formulate cross-entropy based speaker selection to inform relevance inaddition 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 [15, 16].

Let Pr_S be the probability distribution of speakers in the mixed dataset, while Pr_T be the accent-specific distribution of speakers from the mixed set that belong to the target accent. A binary selection variable $\sigma \epsilon \{0, 1\}$ is used to identify target accent speakers. If $\sigma = 1$ for a speaker s from the mixed set, then s belongs to the target accent. The target accent probability distribution can be written as

$$Pr_T[s] = Pr_S[s|\sigma = 1] \tag{2}$$

By Bayes rule,

$$Pr_S[s] = \frac{Pr_S[s|\sigma = 1]Pr[\sigma = 1]}{Pr[\sigma = 1|s]} = \frac{Pr[\sigma = 1]}{Pr[\sigma = 1|s]} Pr_T[s] \tag{3}$$

The probability that a given speaker s belongs to the target accent $Pr[\sigma = 1|s]$ is then

$$Pr[\sigma = 1|s] = \frac{Pr_T[s]}{Pr_S[s]} Pr[\sigma = 1] \tag{4}$$

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

$$Selection\ Score \approx \log\ Pr[s|\lambda_T] - \log\ Pr[s|\lambda_S] \tag{5}$$

The speakers in our mixed set may have different durations, so we need to remove any correlation of the selection score with the duration. The log-likelihoods can be normalized by the number of frames for each speaker. The length normalized log-likelihood is also the cross-entropy of the data given the model [17, 2] with sign reversed. The final score for target data selection is given by

$$Selection\ Score = (-H_{\lambda_T}[s]) - (-H_{\lambda_S}[s]) \tag{6}$$

where

$$H_\lambda(s) = -\frac{1}{T_s} \sum_{u=1}^{U_s} \sum_{t=1}^{u_T} \log\ p(u_t|\lambda) \tag{7}$$

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 1 shows the pseudo code for cross-entropy based semi-supervised learning for target accent adaptation. We note that in the formulation above, we assumed that the initial model is trained on the audio data with multiple accents. The selection score can still be shown to be valid in the case where the seed ASR is trained on a source accent different from the target accent.

Algorithm 1 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
    $Loglikes_S := Score(\lambda_S, x)$
    $Loglike_T := Score(\lambda_T, x)$
    $Len := Length(x)$
    $Score := (Loglike_T - Loglikes_S)/Len$
    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 $\lambda_T$
```

3. IMPLEMENTATION DETAILS

We start with a GMM-HMM model trained on the source data. We adapt this model to the target accent using a small amount of manually transcribed target data. We use enhanced polyphone decision tree adaptation based on semi-continuous models (SPDTS) [2] 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 [18] 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 [19].

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 6.

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 1 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 7 becomes

$$H_\lambda(s) = -\frac{1}{T_s}\sum_{u=1}^{U_s}\sum_{t=1}^{u_T}\log\ p(u_t|\lambda, W_r) \tag{8}$$

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 [12]. 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

$$H_\lambda(s) = -\frac{1}{T_s}\sum_{u=1}^{U_s}\sum_{t=1}^{u_T}\log\ p(u_t|\lambda, W) \tag{9}$$

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

$$p(u|\lambda, W) = \sum_{w=1}^{W} p(u|\lambda, w)p(w) \tag{10}$$

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.

4. EXPERIMENTS

4.1. Datasets

We conducted experiments on Arabic and English accented datasets derived from GALE and WSJ corpora. For Arabic, we used Modern Standard Arabic (MSA) as the source accent and Levantine Arabic as the target. 1100 hours of Broadcast News (BN) portion of GALE database is used as the *Training set*. This is not purely MSA data but may also contain limited amount of other accents. We randomly selected 10 hours from 30 hours of Broadcast Conversations (BC) marked as Levantine by LDC as our *Adaptation set*. The remaining 20 hours of Levantine is combined with 200 hours of other BC data to form the *Mixed set*. For English, we used American accent (WSJ database) as the source accent and British English (WSJCAM0 database) as the target. Our supervised adaptation setup has 66 hours of source accent as *Training set* and 3 hours of target as *Adaptation set*. We combined the remaining 12 hours of target from WSJCAM0 and 15 hours of source data from WSJ0 corpora as our *Mixed set*. The test sets, LM, dictionary are same as the setup followed in [2]. Table 1 summarizes the datasets used and their statistics.

Table 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

4.2. Baselines and Setup

Our baselines are GMM-HMM based fully-continuous systems with LDA, Semi-Tied Covariance (STC) and Speaker-Adaptive (SA) Maximum Likelihood (ML) training. More details about these systems including front-end, training procedure, etc. are reported in [20, 2]. 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 2 shows the Word-Error Rates (WER) of the baseline and adapted systems.

Table 2. *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

4.3. Semi-supervised Learning

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.

4.3.1. 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.

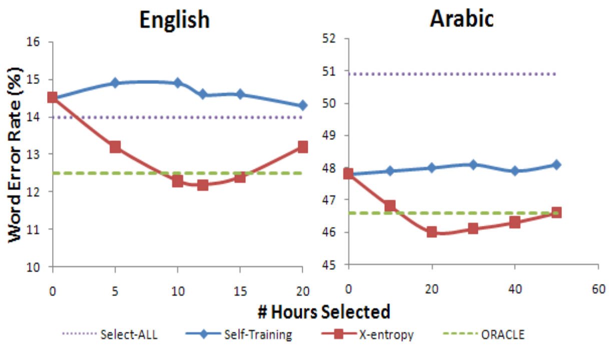

Fig. 1. *Semi-supervised data selection with transcriptions*

In both Arabic and English, self-training does not produce any improvements from semi-supervised learning over the supervised adaptation baseline. In Table.2, the WER on the target test set is higher than the source test set, even for the adapted systems. Hence, log-likelihood or confidence based data selection based on the adapted model cannot differentiate between relevant data (target accent) and irrelevant data (source accent). The initial speakers selected for self-training belong exclusively to the source accent which is the reason for the poor performance of re-trained models. This experiment clearly shows that data selection based only on confidence scores fails when the source ASR is adapted on a limited target data and the unlabeled data is not homogeneous. Cross-entropy based selection on the other hand, relies on change in log-likelihood before and after adaptation to identify the relevant speakers from the mixed set. It obtains an improvement of 2.3% absolute (or 15.9% relative @12 hours) for English and 1.8% absolute (or 3.8% relative @20 hours) for Arabic over the supervised baseline.

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

4.3.2. 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. 1. For cross-entropy based data selection, speaker level selection is based on the difference in lattice likelihoods as in Eq 9. 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.

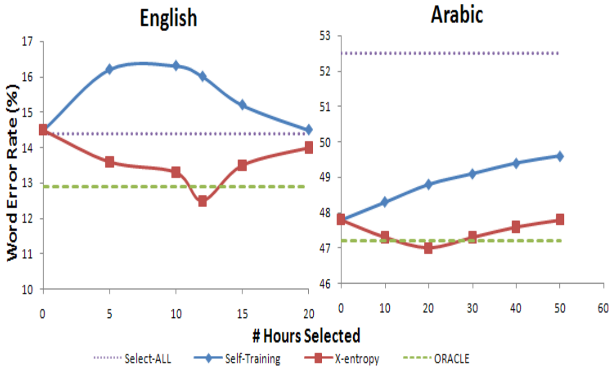

Fig. 2. *Semi-supervised data selection without transcriptions*

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 [1].

5. CONCLUSION AND FUTURE WORK

We presented cross-entropy based semi-supervised learning to select additional training data for the goal of accent adaptation. We evaluated the technique against self-training on two Arabic and English tasks and showed that self-training failed to obtain any improvements by using additional data over the supervised baseline. Our cross-entropy based data selection method successfully identifies relevant data with or without transcriptions, to obtain further 2.0-15.9% relative improvement over the supervised baseline adapted on a limited target dataset. We plan to extend this semi-supervised learning framework to include uncertainity based discriminative training as proposed in [21]. We also plan to extend this framework to active learning [13, 22, 23, 24] for accent adaptation to select appropriate subset for human annotation for further improvements.

6. REFERENCES

[1] Hagen Soltau, Lidia Mangu, and Fadi Biadsy, "From modern standard arabic to levantine asr: Leveraging gale for dialects," in *ASRU*, 2011, pp. 266–271.

[2] Udhyakumar Nallasamy, Florian Metze, and Tanja Schultz, "Enhanced polyphone decision tree adaptation for accented speech recognition," in *INTERSPEECH*, 2012.

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

[4] Kai Yu, Mark J. F. Gales, Lan Wang, and Philip C. Woodland, "Unsupervised training and directed manual transcription for LVCSR," *Speech Communication*, vol. 52, no. 7-8, pp. 652–663, 2010.

[5] Frank Wessel and Hermann Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 23–31, 2005.

[6] Thomas Kemp and Alex Waibel, "Unsupervised training of a speech recognizer: recent experiments," in *EUROSPEECH*, 1999.

[7] Bhuvana Ramabhadran, "Exploiting large quantities of spontaneous speech for unsupervised training of acoustic models," in *INTERSPEECH*, 2005, pp. 1617–1620.

[8] Jeff Z. Ma and Richard M. Schwartz, "Unsupervised versus supervised training of acoustic models," in *INTERSPEECH*, 2008, pp. 2374–2377.

[9] 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*, 2009, pp. 88–91.

[10] Scott Novotney, Richard M. Schwartz, and Sanjeev Khudanpur, "Unsupervised arabic dialect adaptation with self-training," in *INTERSPEECH*, 2011, pp. 541–544.

[11] Wikipedia, "Dialect," `http://en.wikipedia.org/wiki/Dialect`.

[12] Lidia Mangu, Eric Brill, and Andreas Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 373–400, 2000.

[13] 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*, vol. 24, no. 3, pp. 433–444, 2010.

[14] Thiago Fraga-Silva, Jean-Luc Gauvain, and Lori Lamel, "Lattice-based unsupervised acoustic model training," in *ICASSP*, 2011, pp. 4656–4659.

[15] John Blitzer and Hal Daumé III, "ICML tutorial on domain adaptation," `http://adaptationtutorial.blitzer.com`, June 2010.

[16] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh, "Sample selection bias correction theory," in *ALT*, 2008, pp. 38–53.

[17] Robert C. Moore and William Lewis, "Intelligent selection of language model training data," in *ACL (Short Papers)*, 2010, pp. 220–224.

[18] 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*, vol. 2, no. 2, pp. 291–298, 1994.

[19] Daniel Povey and Kaisheng Yao, "A basis representation of constrained mllr transforms for robust adaptation," *Computer Speech & Language*, vol. 26, no. 1, pp. 35–51, 2012.

[20] Florian Metze, Roger Hsiao, Qin Jin, Udhyakumar Nallasamy, and Tanja Schultz, "The 2010 cmu gale speech-to-text system," in *INTERSPEECH*, 2010, pp. 1501–1504.

[21] Xiaodong Cui, Jing Huang, and Jen-Tzung Chien, "Multi-view and multi-objective semi-supervised learning for hmm-based automatic speech recognition," *IEEE Transactions on Audio, Speech and Language processing*, vol. 20, no. 7, pp. 1923–1935, 2012.

[22] Avishek Saha, Piyush Rai, Hal Daumé III, Suresh Venkatasubramanian, and Scott L. DuVall, "Active supervised domain adaptation," in *ECML/PKDD (3)*, 2011, pp. 97–112.

[23] Giuseppe Riccardi and Dilek Hakkani-Tür, "Active learning: theory and applications to automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 504–511, 2005.

[24] N. Itoh, T.N. Sainath, D.N. Jiang, J. Zhou, and B. Ramabhadran, "N-best entropy based data selection for acoustic modeling," in *ICASSP*, 2012, pp. 4133–4136.