NEIGHBOUR SELECTION AND ADAPTATION FOR RAPID SPEAKER-DEPENDENT ASR

Udhyakumar Nallasamy[1]*, Mark Fuhs*[2]*, Monika Woszczyna*[2]*, Florian Metze*[1] *and Tanja Schultz*[1,3]

[1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
[2]Speech Technology Group, M*Modal Inc., Pittsburgh, USA
[3]Cognitive Systems Labs, Karlsruhe Institute of Technology, Germany

{unallasa, fmetze}@cs.cmu.edu, {mark.fuhs, monika.woszczyna}@mmodal.com, tanja.schultz@kit.edu

ABSTRACT

Speaker dependent (SD) ASR systems have significantly lower word error rates (WER) compared to speaker independent (SI) systems. However, SD systems require sufficient training data from the target speaker, which is impractical to collect in a short time. We present a technique for training SD models using just few minutes of speaker's data. We compensate for the lack of adequate speaker-specific data by selecting neighbours from a database of existing speakers who are acoustically close to the target speaker. These neighbours provide ample training data, which is used to adapt the SI model to obtain an initial SD model for the new speaker with significantly lower WER. We evaluate various neighbour selection algorithms on a large-scale medical transcription task and report significant reduction in WER using only 5 mins of speaker-specific data. We conduct a detailed analysis of various factors such as gender and accent in the neighbour selection. Finally, we study neighbour selection and adaptation in the context of discriminative objective function.

Index Terms— Speech recognition, acoustic modeling, speaker adaptation, data selection approaches

1. INTRODUCTION

Speaker specific characteristics such as age, gender, vocal-tract length and accent have significant influence on the ASR acoustic models. Speaker-independent systems trained by pooling data from wide range of speakers perform poorly for speakers whose features are under-represented in the training set, e.g. non-natives, female speakers, etc. Speaker-adaptive systems handle these variations using various normalization [1] and adaptation techniques [2]. Speaker-dependent ASR systems on the other hand, address this issue by training models solely on data specific to the target speaker. While SD systems perform significantly better compared to an SI system [3], they require fairly large amounts of the target speaker's training data, which is time-consuming to collect.

In a real-world task such as dictation, the user starts off with an SI system out of the box and the system adapts to the speaker's data with continued usage. Unfortunately, the new speaker has to painfully navigate through this adaptation phase with a low accuracy SI system until he/she has produced sufficient data for adapting the initial model. This issue is more serious for accented speakers who encounter significantly higher word error rates with SI system compared to native speakers. Customer satisfaction and adoption rates for commercial ASR systems suffer significantly throughout this transition period. In this paper, we aim to shorten this adaptation period by creating better SD models as soon as the initial ASR acquires just a few minutes of data from the new user.

We address the challenge of building SD models by automatically selecting acoustically similar speakers to the target speaker, or *neighbours* from a large and diverse set of existing users with large amounts of training data. We use a few minutes of the speaker's data to select the neighbours, so the adaptation can be performed sooner than waiting for sufficient data from the user. We utilize the neighbours' data to build an initial SD system for the target speaker. We show that such a neighbours initialized SD system performs significantly better compared to the baseline SI models, thus helping to reduce the adaptation interval for the new speaker.

2. RELATED WORK

Speaker adaptation has a long history in ASR with popular techniques such as Maximum A-Posteriori (MAP) adaptation [4], Maximum Likelihood Linear Regression (MLLR) [5] and Constrained-MLLR (CMLLR) [2]. However, these techniques are confined to the available adaptation data, which is only a few minutes in our case. To address the issue of limited data, several approaches have been previously proposed. They can be classified into 3 groups. In eigen-based techniques, a low-dimensional projection of model [6, 7] or transform parameters [8, 9] is used to reliably estimate parameters with less data. These techniques are most effective with adaptation data of a few seconds and saturate to the performance of regular MLLR/CMLLR with data more than 10 seconds.

In clustering based approaches, the training dataset is clustered into multiple groups with individual [10, 11] or shared set of models [12]. The test speaker is assigned to one or more of the clusters and uses the models estimated on their respective speaker groups. These techniques are computationally efficient as training the clusters can be done offline and only the cluster assignment is carried out during decoding. However, it is sub-optimal to precluster the speakers as its difficult to obtain representative clusters for different factors such as age, gender, accent, etc. that influence the acoustic space of each speaker.

Rank-and-select approaches attempt to find a ranked list of acoustically relevant neighbours specific to each test speaker [13, 14, 15, 16] and uses the neighbours' data for adaptation. It can also be viewed as an instance of exemplar-based technique [17] at the speaker level. Most of the previous attempts used some form of approaximation of SI and SD models to select the neighbours based on the acoustic match. [18] used single gaussian models to train SD models on few secs of training data. [15, 16] used gaussian clustering and linear transformation to select source speakers for augmenting the adaptation data. In this paper, we use the existing

First author is an intern at M*Modal Inc. during the course of this work

ASR models to directly compute the likelihood of target adaptation data. We explore several variations of this technique and empirically evaluate them based on their performance on the target speaker test set. We also investigate different parameters involved in the selection including the number of neighbours, size of adaptation data, etc.

3. NEIGHBOUR SELECTION TECHNIQUES

We study two different neighbour selection techniques in our initial experiments - likelihood based and transformation based. The likelihood based approach aims to find source speakers in the training set who are *close* to the given target speaker. It is performed using the following steps:

- The SI model is adapted to each of the source speakers in the database.
- The resulting source SD models are used to calculate the likelihood of the target speaker's data. Given a source model λ_S for the source speaker S, adaptation utterances U_T and their reference transcriptions W_r for the target speaker T, the likelihood is calculated as

$$\text{Likelihood}_T(S) = \sum_{u \in U_T} logP(O_u, W_r|\lambda_S) \qquad (1)$$

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

In transformation based approach, source SD models are computed as the first step similar to the likelihood based technique. We do an additional step of adapting the SD models on the target speaker's data before calculating the likelihood score. The likelihood in transformation based approach is given by

$$\text{Likelihood}_T(S) = \sum_{u \in U_T} logP(O_u, W_r|f_T(\lambda_S)) \qquad (2)$$

where $f_T(\lambda_S)$ is the source model adapted on the target speaker's data. In our case, we use a regression-tree based MLLR for the transformation function f_T. The source speakers are ranked as before for the selection. The transformation based neighbour selection attempts to choose neighbours who can be *transformed* into the target speaker. The extra adaptation step compensates for any mismatch between the source and target speakers, that can be modeled by linear transformations, e.g. channel variations. Section 7 explores neighbour selection and adaptation using discriminative objective function.

Once the neighbours are chosen, we adapt the SI model using data from the selected neighbours. The neighbour adapted model is used to initialize the SD system for the new speaker. As we get more of the target speaker's data, we continue to adapt the initial model to obtain an accurate SD model for the speaker.

4. EXPERIMENTS

4.1. Database and setup

We conduct our experiments on an 8kHz, telephony quality, English medical transcription task. Table 1 lists the different datasets used in our experiments. Medical dictation is a fast-paced speaking style compared to typical conversational speech. Our training dataset contains medical reports dictated from 1878 training speakers with a maximum of 1 hour per speaker. The database has speakers with different accents, varying telephony channels and background noise levels. The total size of the dataset is 1450 hours. A set of 10 South-Asian accented speakers, independent of the training set form our target speakers. We use 5 minutes for each target speaker as our development set. For the test set, the same speakers with 1 hour of speech are used. As mentioned before, we are interested in the accuracy of SD models after 5 minutes of adaptation. Hence, we don't perform any second pass, unsupervised adaptation on the 1 hour test set. We report word error rates averaged across the 10 target speakers.

Table 1. *Datasets and their statistics.*

Dataset	Speakers	#Hours	&words
Train	1878	1450	1.2M
Dev	10	0.83	8.3K
Test	10	10.72	86K

4.2. Baseline system

The SI system is a fully-continuous, ML trained, GMM-HMM based ASR using 3000 context-dependent states and 86K gaussians. The system uses MFCC features, Vocal Tract Length normalization (VTLN) and a global Semi-tied Covariance (STC) matrix trained using ML criterion. The decoder uses a 4-gram language model with a vocabulary size of 53K words. The language model has a OOV of 0.8% on the test set. As a first step, the SI model is adapted on 5mins of the development set using regression-tree based MLLR [19]. The number of tranforms for MLLR is automatically selected based on the amount of available adaptation data. In our case, we ended up with an average of 10 MLLR transforms given 5mins of adaptation data for each target speaker.

Table 2 shows the WER of SI and MLLR adapted systems. The MLLR adapted system produces a relative improvement of 10.4% over the SI model. Additional improvements can be obtained by training canonical models using SAT and CMLLR. However, the CMLLR matrices for the test speakers have to be computed on the adaptation data as this is a one-pass dictation system. Such a SAT setup didn't give us any significant improvement on top of regression-tree based MLLR adaptation with 5 mins of speaker-specific data in our previous experiments, so we didnt include SAT in our baseline.

Table 2. *Baseline WERs.*

System	Test set	WER
SI	South Asian	45.73
SI + MLLR	South Asian	40.99
SI	Native	29.89

To put the WER on accented speakers in context, we also include the WER of the SI system on a test set of 15 native US English speakers from the same task. It shows that a new South Asian accented speaker will start with a significantly worse (53% relative) ASR for dictation compared to a native US English speaker, before any adaptation. However, we note that the South Asian accent is not a homogeneous group of speakers. The test set statistics and WER of the SI model broken down by individual speakers is listed in Table 3. It shows the wide differences between the South Asian speakers in the test set. The SI WERs vary anywhere from 19.7% to 63.8%.

Hence, it is important to select neighbours to match each speaker's individual characteristics, for adaptation.

Table 3. *SI WERs for South-Asian speakers.*

Speaker	Test Data (15 Reports)		Test WER (%)
	Words	#Hours	
1	15673	2.28	19.7
2	7690	0.89	30.5
3	7702	0.80	42.7
4	8837	1.11	37.3
5	6762	0.98	48.0
6	8221	1.02	58.5
7	12301	1.55	63.8
8	6761	0.73	59.7
9	4535	0.45	53.5
10	7782	0.91	43.6
Avg	**8626.4**	**1.07**	**45.73**

4.3. Neighbour selection

In likelihood based selection, we adapt the SI model to the source speaker using MAP adaptation. We then compute the likelihood of the target speaker's data (5 mins) on the adapted model. The source speakers are ranked based on the likelihood score. In the transformation based technique, we compute an additional regression-tree based MLLR for the source model on the target data before the likelihood computation. We select 20 neighbours using each criteria. We constrained the neighbours to have at least 15 minutes of speech to ensure sufficient data for adaptation. Once the neighbours are selected, we use MAP to adapt the SI model on the neighbours' data. The neighbour initialized model is further adapted using MLLR on the target speaker's data. The final SD models are used to decode the test set. Table 4 shows the WER for the likelihood and transformation based selection. The results show that transformation based neighbour selection outperforms the likelihood based approach. It also has 26.3% relative lower WER than the SI and 17.8% relative lower than MLLR adapted baseline.

Our setup of creating SD models for each source speaker might seem computationally demanding. However, the neighbours are chosen from a set of existing speakers with large amounts of data. These speakers already have SD systems trained for their own dictation. We only need to access their parameters instead of creating source SD models from scratch during selection. Once the neighbours are chosen, the data for these speakers are accessed for adapting the SI model. Table 5 shows speaker-wise WER for the SI, SI + MLLR and neighbour MAP + MLLR systems. The improvements over SI+MLLR are between 10.9% and 31.2% on this test set. This shows that neighbour selection produces improvements for speakers over a wide range of WERs. The following sections will analyze varying these parameters and their influence on target WER. All the experiments hereforth will use transformation based neighbour selection.

4.3.1. Varying the number of neighbours

In this section, we vary the number of neighbours chosen, from 1 to 80 and use them to initialize the SD system for the target speaker. Table 6 lists the WERs of adapting with varying the neighbours. The adaptation step after neighbour selection, involves MAP adaptation on the neighbour data followed by MLLR adaptation on the target data. We list both the WERs below.

Table 4. *WER for neighbour selection techniques.*

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

Table 5. *Adaptation WERs for South-Asian speakers.*

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

Table 6. *Analysis of varying number of neighbours.*

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

From Table 6, it is clear that 20 neighbours produces the lowest WER. However, neighbours 5 and 10 are very close to the WER of 20 neighbours.

4.3.2. Varying the amount of adaptation data

We vary the amount of target adaptation data to measure its effect on the neighbour selection. We should note here that, to select different amounts of target speaker data, we start adding utterances of each speaker to the development set until we reach the desired time in minutes. However, we do not excise utterances to meet the time limit, so the exact duration will be slightly higher than the expected length. We vary the target data by 1, 2, 5 and 60 minutes for neighbour selection. In each case, we choose 20 neighbours and perform neighbour-MAP followed by target-MLLR adaptation. Table 7 lists the WERs for neighbour selection carried out for different amounts of development data. To add clarity, we list the exact amount of target adaptation data (averaged across speakers) chosen for each case.

Neighbour MAP WER in Table 7 can be used to compare the

Table 7. *Analysis of varying adaptation data.*

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

speaker selection across different amount of adaptation data. The results show several interesting properties. Focusing on the first three rows, we get better neighbours with increasing target data. However, the neighbours chosen with just 2.68 minutes perform quite close to the ones selected with 5.82 minutes of target speaker data. This is attractive as most dictation systems perform an enrollment step which guides the new user to read out a few phonetically balanced sentences. The average amount of data collected during the enrollment step is around 2 minutes which could be used to select the neighbours, rather than waiting for the speaker to start using the system.

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

5. ANALYSIS

5.1. Influence of gender and accent

In this section we analyze the neighbours selected to study the influence of various factors such as gender and accent. Our training data is manually annotated with accent and gender labels. For a few speakers without gender labels, we assign them based on their VTLN [1] warpfactors. We use the annotations to measure the influence of different factors on the neighbours selected for a target speaker. We note that 99% of the neighbours selected match the gender of the reference speaker. Hence, we can conclude that gender has a decisive impact in the neighbour selection. Figure 1 shows the cumulative count of South-Asian and non-South-Asian neighbours in each rank added across all target speakers.

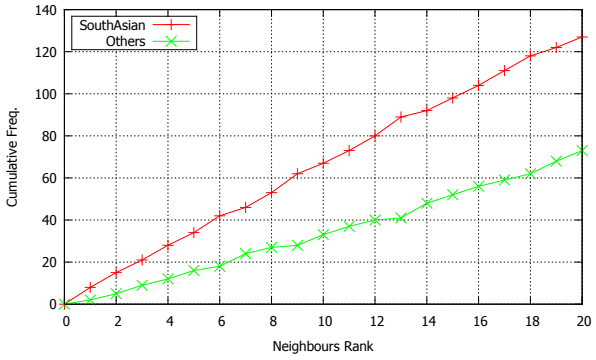

Fig. 1. *Cumulative frequency of neighbours at each rank.*

The graph clearly shows that South-Asian speakers are ranked higher than others in the neighbours list. We conduct Mann-Whitney U test [20], a non-parametric rank test to verify the influence of accent. We consider 100 ranked neighbours selected for each speaker and group them into South-Asian and non-South-Asian categories. The test showed significant difference ($p < 0.001$) between the ranks of the two groups, thus confirming accent has significant influence on choosing neighbours.

5.2. Automatic selection vs. manual annotations

In this section, we conduct experiments to compare automatic selection with choosing neighbours based on the manual annotations. We have 168 South-Asian speakers labeled in our training set. We use the gender and accent labels to explicitly choose neighbours that match the target speaker to compare it with our automatic selection technique. In all cases, once the neighbours are decided we perform MAP adaptation on neighbours' data and MLLR on the target speaker's speech. Table 8 shows the WERs of adapted systems on automatically selected neighbours and the ones based on manual labels.

Table 8. *Automatic selection Vs. Manual annotations.*

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

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

In the second set of experiments, we combine automatic selection with manual annotations, by running transformation based neighbour search on the accent subset instead of the whole training set. Table 9 lists the WER of automatic selection without and with manual annotations.

Table 9. *Automatic selection using manual annotations.*

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

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

6. VARYING TARGET SPEAKER'S DATA

In this section, we examine the behaviour of SI and neighbours initialized SD models with increasing adaptation data. For each datapoint, we adapt both systems on the available adaptation data and report WER on the test set. We calculate MLLR on the target data and use the transformed means as a prior model for the ensuing MAP adaptation. The combined adaptation performed better than using MLLR or MAP alone. Figure 2 plots the WER for both systems. The datapoint at zero SD data, refers to the SI baseline.

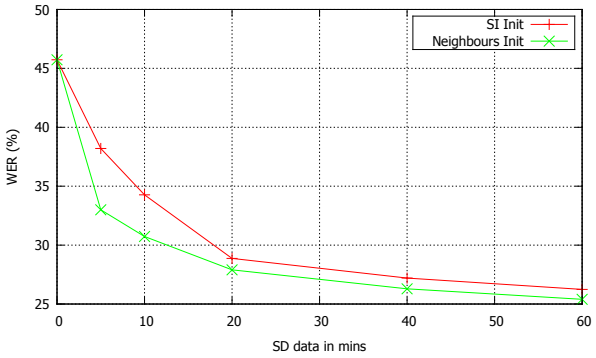

Fig. 2. *WER for SouthAsian speakers*

It is interesting to note that, although the neighbours are chosen with only 5 minutes of target speech, the neighbours initialized system continues to perform better than SI with increased adaptation data. To understand the impact of the neighbour adaptation technique on native speakers, we conducted the same experiment on the test set of 15 US English speakers. As in the South-Asian case, we used 5 minutes of each speaker to select the neighbours. Figure 3 shows the WER plot of SI-Init and Neighbours-Init systems with increasing adaptation data. We find the same pattern for both the systems as with non-native speakers. However, the total improvement is less (7% relative at 5 minutes) compared to the South-Asian case (15% relative at 5 minutes), which is expected as the majority of training set data consist of native speakers.

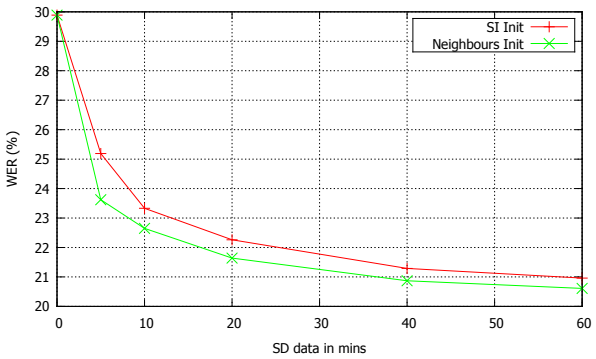

Fig. 3. *WER for Native speakers*

7. DISCRIMINATIVE SELECTION AND ADAPTATION

The experiments so far have been conducted with ML trained SI model. In this section, we investigate the neighbour selection and adaptation using discriminatively trained (DT) models. Our DT SI model is trained using state-level Minimum Bayes Risk (sMBR) [21] for 8 iterations with an I-smoothing weight τ of 350. It obtains a WER of 34.55% which is 24.4% relatively lower than ML SI model. Our first step involved plugging this model in tranformation based speaker selection approach to choose 20 neighbours. Table 10 shows the WER for different selection and adaptation setups with SI model in row 1. It can be seen that there is almost no improvement by adapting the DT SI model on the neighbours data using ML-MAP criterion. We implemented discriminative adaptation [22] of the SI model on the neighbours data, which yielded a relative improvement of 8.5% compared to the unadapted model.

We also investigated using discriminative criterion in neighbour selection, in addition to adaptation. In this approach, the adaptation data of 5 mins is decoded using the SI model to create lattices for the target speaker. We choose source speakers whose models maximize the sMBR accuracy on the target speaker lattices. The sMBR accuracy is calculated as

$$\text{sMBR Accuracy}_T(S) = \sum_{u \in U_T} \sum_{W \in W'} \gamma(O_u, W | f_T(\lambda_S)) A(W_r, W) \quad (3)$$

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

Table 10. *Discriminative selection and adaptation.*

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

From Table 10 row 4, it is interesting to note that neighbours selected using sMBR accuracy produce 7.5% relative improvement over SI model, using ML MAP adaptation. Comparing rows 2 and 4, we note that discriminative selection can lead to neighbours who produce less WER on target speaker data than ML based selection. However, the gains from discriminative selection and adaptation are not additive but the combined technique shown in row 5, still produces the best result of 10.3% relative improvement over the unadapted system. Figure 4 shows the WER of SI and neighbours-initialized systems with increasing target data. We performed ML adaptation in each bin as we didnt get any additional improvement through DT adaptation on the relatively small amount of available speaker-specific data (<= 1 hour)

8. CONCLUSION

We have presented an adaptation technique to build SD models with a few minutes of target speaker's data. We obtained an improvement of 23% relative over SI with just 5 minutes of the target speakers' data. We analyzed the selected neighbours and showed that accent and gender play a crucial role in their selection. We also compared the automatic selection with choosing neighbours based on manual

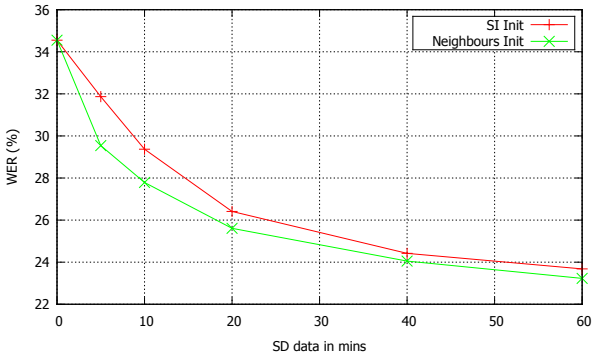

Fig. 4. *WER for Discriminative selection and adaptation*

annotations and concluded that the automatic approach performed better. Finally, we studied neighbour selection and adaptation with discriminative objective functions and showed that they perform better than ML based alternatives. As part of the future work we plan to investigate Deep Neural Networks [23] and Bottle-neck features [24] in the context of SD models and neighbours' adaptation. We would also like to extend the work to unsupervised adaptation and evaluate the benefit of neighbour selection in that scenario.

9. ACKNOWLEDGEMENTS

We thank Thomas Schaaf and members of M*Modal speech technology group for insightful discussions.

10. REFERENCES

[1] Puming Zhan and Alex Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition," Tech. Rep., Carnegie Mellon University, 1997.

[2] M. J. F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75–98, 1998.

[3] Yu Shi and Eric Chang, "Studies in massively speaker-specific speech recognition," in *INTERSPEECH*, 2002.

[4] 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.

[5] C. J. Leggetter and Philip C. Woodland, "Speaker adaptation of continuous density hmms using multivariate linear regression," in *ICSLP*, 1994.

[6] Roland Kuhn, Patrick Nguyen, Jean-Claude Junqua, Lloyd Goldwasser, Nancy Niedzielski, Steven Fincke, Ken Field, and Matteo Contolini, "Eigenvoices for speaker adaptation," in *ICSLP*, 1998.

[7] Michiel Bacchiani, "Rapid adaptation for mobile speech applications," in *ICASSP*, 2013.

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

[9] 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.

[10] Ananth Sankar, Françoise Beaufays, and Vassilios Digalakis, "Training data clustering for improved speech recognition," in *EUROSPEECH*, 1995.

[11] Françoise Beaufays, Vincent Vanhoucke, and Brian Strope, "Unsupervised discovery and training of maximally dissimilar cluster models," in *INTERSPEECH*, 2010, pp. 66–69.

[12] Mark J. F. Gales, "Cluster adaptive training of hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 417–428, 2000.

[13] Mukund Padmanabhan, Lalit R. Bahl, David Nahamoo, and Michael A. Picheny, "Speaker clustering and transformation for speaker adaptation in speech recognition systems," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 71–77, 1998.

[14] Chao Huang, Tao Chen, and Eric Chang, "Speaker selection training for large vocabulary continuous speech recognition," in *ICASSP*, 2002, pp. 609–612.

[15] Jian Wu and Eric Chang, "Cohorts based custom models for rapid speaker and dialect adaptation," in *INTERSPEECH*, 2001, pp. 1261–1264.

[16] Ravichander Vipperla, Steve Renals, and Joe Frankel, "Augmentation of adaptation data," in *INTERSPEECH*, 2010, pp. 530–533.

[17] Tara N. Sainath, Bhuvana Ramabhadran, David Nahamoo, Dimitri Kanevsky, Dirk Van Compernolle, Kris Demuynck, Jort F. Gemmeke, Jerome R. Bellegarda, and Shiva Sundaram, "Exemplar-based processing for speech recognition: An overview," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 98–113, 2012.

[18] Jing Huang and Mukund Padmanabhan, "A study of adaptation techniques on a voicemail transcription task," in *EUROSPEECH*, 1999.

[19] M. J. F. Gales, "The generation and use of regression class trees for MLLR adaptation," Tech. Rep., Cambridge University, 1996.

[20] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.

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

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

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

[24] Tara N. Sainath, Brian Kingsbury, and Bhuvana Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," in *ICASSP*, 2012, pp. 4153–4156.