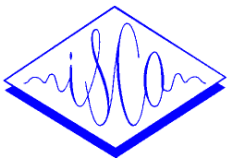

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

EFFICIENTLY USING SPEAKER ADAPTATION DATA

Chengyi Zheng[1] *and Yonghong Yan*[2]
[1]Center for Spoken Language Understanding , Oregon Graduate Institute
[2] Intel Corporation, USA

ABSTRACT

Transformation based speaker adaptation techniques, such as Maximum Likelihood Linear Regression (MLLR) [1] require a large amount of adaptation data to robustly estimate the transform matrices. In this paper, we present a new adaptation scheme that adjusts the adaptation data according to the feedback from recognizer. By giving different weights to different parts of the adaptation data, the proposed scheme can make use of the adaptation data more efficiently. Experiments on the WSJ 20K task show that this method achieved an additional 10% relative word error rate reduction in supervised adaptation and 2% reduction in unsupervised adaptation compared with conventional MLLR approach.

1. INTRODUCTION

Speaker adaptation was developed for speech recognition systems that face multiple speakers. Speaker adaptation was used to adapt the Speaker Independent (SI) acoustic model to the Speaker Dependent (SD) model with a small amount of speech data from a target speaker. Speaker adaptation techniques can be used in supervised and unsupervised mode. In supervised mode, the correct transcription is known, while in unsupervised mode, no correct transcription is available. Unsupervised adaptation uses the best available model to generate the transcription of the adaptation data.

One challenge to speaker adaptation is the limited amount of adaptation data. How to efficiently use these data is thus crucial. In the conventional adaptation scheme, all the adaptation data are treated equally, that is, no part of the adaptation data is more important than others. Due to the non-uniform distribution (over all the acoustic models) of the adaptation data, some acoustic models might be under-adapted. The words, whose pronunciations are represented by these under-adapted models, are likely to be mis-recognized.

We propose an adaptation scheme to make use of the adaptation data more efficiently by obtaining feedback from the decoding. In supervised mode, we focus on emphasizing the mis-recognized words, while in unsupervised mode we focus on locating and removing the near miss words from the adaptation data set.

In Section 2 we sketch a very simple self-adjustable adaptation scheme based on the conventional adaptation schemes. In Section 3 and 4 we describe some experiments on the supervised and unsupervised speaker adaptation. Finally we summarize our major findings and outline our future work.

2. SELF-ADJUSTABLE SPEAKER ADAPTATION

The conventional adaptation scheme is as follows:

1. Given some adaptation enrollment data and a SI model, collect statistics on the enrollment data and perform speaker adaptation on the SI model.
2. Decoding the test utterances with the adapted acoustic model.

Such a scheme uses the enrollment data only once and does not incorporate any feedback from decoding. It is fast in practice. However for certain applications where performance is more important, we propose the following iterative adaptation scheme, as illustrated in Figure 1, that dynamically adjusts enrollment data to incorporate feedback from decoding on the enrollment data.

1. Denote $\boldsymbol{M}$ the initial **SI** model, and $\boldsymbol{A}$ the enrollment data set.
2. Perform speech recognition on data set $\boldsymbol{A}$ based on model $\boldsymbol{M}$.
3. Adjust $\boldsymbol{A}$ to $\boldsymbol{A}'$ according to the decoding results from step 2. Emphasizing or de-emphasizing certain parts of $\boldsymbol{A}$ with weights based on these results.
4. Adapt model $\boldsymbol{M}$ to $\boldsymbol{M}'$ using enrollment data $\boldsymbol{A}'$
5. Repeat step 2 and 3 with the updated $\boldsymbol{M}'$

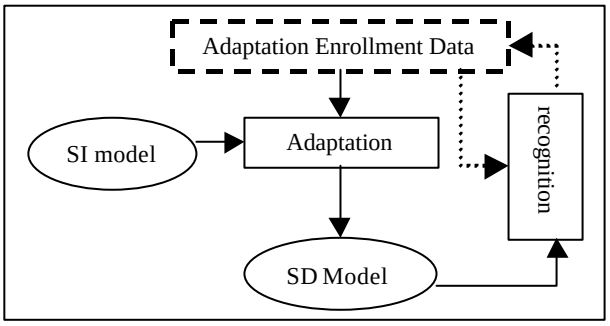

Figure 1: Block diagram of the modified Speaker Adaptation (Bold-dashed lines refer to the modified part)

This modified adaptation scheme has the following advantages:

1. This scheme provides a way to incorporate the recognition feedback. For those words (or utterances in the enrollment data set) that are not well learnt by adaptation, they might be better adapted with the given higher weights.
2. When only limited amounts of enrollment data are available, this scheme of iterative bootstrapping makes better use of that limited data.
3. The scheme can be extended to the unsupervised adaptation where references may contain errors.

We will present some methods based on this scheme in the following sections.

3. SUPERVISED SPEAKER ADAPTATION

Throughout our experiments in this paper, a crossword triphone decoder is used as the baseline system [2]. The acoustic front-end extracts 12 mel-frequency cepstral coefficients (MFCC), energy, and their first and second derivatives (total of 39 parameters) from every 10ms frame. Cepstrum Mean Normalization is performed on each utterance.

The acoustic model is trained on the standard SI 284 training set, and is state-tied through the decision tree algorithm. The standard EM algorithm is used. The final model has a total of 7500 states with 12 mixture components per state.

We perform the supervised tests on a self-constructed test set based on the WSJ 20K Nov92 test set, which has 8 speakers and 41 utterances per speaker on average. The first 30 sentences of each speaker are used as the enrollment data and the rest as the test data.

The baseline achieved 12.6% word error rate (WER). The standard MLLR adaptation brought the WER down to 9.8%. All the experiments presented in this paper only repeat one of the iterations shown in Figure 1. Further improvements are possible when the number of iterations is greater.

In the experiments, the following hypotheses are to be verified:

1. Weighting more to those words or utterances that were mis-recognized yields better adaptation.

2. Weighting more to the correctly recognized words or utterances does not help the performance.

3. The adaptation performance can be further improved by removing non-acoustic factors, such as OOV (Out Of Vocabulary) words and homonyms.

2.1 Sentence Level Weight Adjustment

We first test the above hypotheses on the sentence level. Following the scheme in Figure 1, the enrollment data was adapted and recognized. Those utterances that are not 100% correctly recognized were picked out as set E and the rest of the sentences as set C. We emphasize E or C by repeating those utterances in set E or C a certain number of times.

Emphasizing set C results in a 12.2% WER *increase*. On the other hand, emphasizing E gives 2% WER reduction as shown in Table 1.

By removing those error utterances merely caused by OOV words from set E, another 5.1% WER reduction was achieved. These experimental results show that removing those mis-recognition utterances that caused by language-model helps to reduce the distortions in set E.

	Average WER	Δ_{WER}
MLLR(baseline)	9.8%	
Weighting Correct	11.0%	+12.2%
Weighting Error	9.6%	-2%
Weighting Error after removing OOV	9.3%	-5.1%

Table 1. Comparison of Sentence Level Supervised Adaptation

2.2 Word Level Weight Adjustment

The similar idea can be extended at the word level, considering each utterance in set E has both correctly recognized words and mis-recognized words. We can further improve the adaptation performance by removing correct words from set E.

For example, in the following pair of reference and hypothesis sentences, only one word '*pilot*' was mis-recognized.

- Reference: the *pilot* an American survived.
- Hypothesis: the *pilots* an American survived.

We want to assign higher weight to word 'pilot' but the whole sentence.

Figure 2 shows the performance curves given different weight values to emphasize the error words in set E. A 9.0% WER was achieved in word level adaptation compared to the 9.3% WER in sentence level adaptation.

Nguyen [3] suggests that adaptation be adjusted according to the N-BEST hypotheses. However, in our case, we reasoned that since lower-ranked hypothesis has less likely to be the correct target, emphasizing error words beyond the first best hypothesis would tend to cause more errors.

A comparison between the 1-best and 5-best systems is shown in Figure 2. The 5-best system achieved 9.3% WER while the 1-best system achieved 9.0% WER. From the figure, we may also see that the 1-best system is less sensitive to different weighting values.

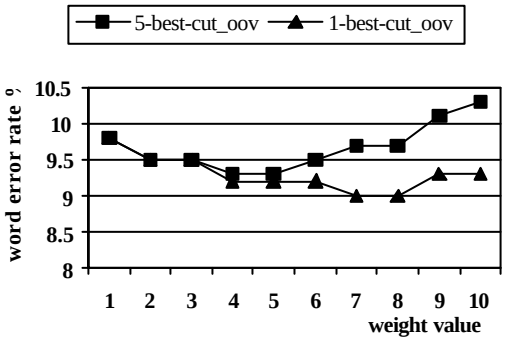

Figure 2: Word Level N-Best supervised adaptation

2.3 Using Likelihood Information to Set Weight

Right now, the weight values are set arbitrarily. It would be better to estimate the weight value from the likelihood information of the references and hypotheses. We propose the following method to calculate the weight value:

1. Run the force alignment program on the enrollment data to get statistics of the references.
2. Decode the enrollment data to get statistics of the 1-best hypothesis.
3. Align the 1-best hypothesis with the reference sentence to obtain the error words.
4. Calculate the average log-likelihood difference per frame according to the following equation:

$$d_{ij} = \frac{\log P(W_r \mid O_j)}{T_r} - \frac{\log P(W_h \mid O_j)}{T_h}, \quad (1)$$

where W_r and T_r are the reference word and its frame count for error word W_j of speaker *i*; W_h and T_h are the 1st best hypothesis word and its frame count correspondingly.

5. Calculate the weight value wgt_i for speaker *i* by averaging L_n over all the error words:

$$wgt_i = \frac{1}{m_i} * \sum_{j=1}^{m_i} \left| d_{ij} \right|, \quad (2)$$

where m_i is the total number of error words of speaker *i.*

Using the above approach for a 1-best hypothesis, an 8.8% word error rate was achieved. (see Table 2)

	Average WER	Δ_{WER}
Baseline (without adaptation)	12.6%	
Conventional adaptation	9.8%	-22%
Sentence level weighting	9.3%	-26%
Word level weighting	9.0%	-29%
Word level weighting using likelihood information	8.8%	-30%

Table 2: Comparison of Supervised Speaker Adaptations

Statistical analyses were performed for each test conditions. Based on a 95% confidence interval ($z_a = 1.96$), the matched pair tests (table 3) show that both the sentence and word level weighting are statistically different with the conventional adaptation. At the same confidence interval, the Wilcoxon tests (table 4) show that the sentence level weighting has no significant different to the conventional adaptation while the word level weightings are statistically different.

Matched-pairs tests	Z Stat	Diff
Sentence level weighting	2.099	yes
Word level weighting	2.385	yes
Word level weighting using likelihood information	3.033	yes

Table 3: Matched-pairs tests results

Wilcoxon tests	Z Stat	Diff
Sentence level weighting	-1.82	No
Word level weighting	-2.10	Yes
Word level weighting using likelihood information	-2.10	Yes

Table 4: Wilcoxon tests results

3. UNSUPERVISED SPEAKER ADAPTATION

Extending the adaptation scheme into unsupervised adaptation raises differences:

- Mis-recognized words mislead the adaptation
- The correctness of the transcription is uncertain. However, it might be able to determine that some words are more likely to be the mis-recognized ones.
- Even if an error word in the hypothesis can be located, there is no particular way to adapt it since the correct reference is unknown.

Therefore, in unsupervised mode, the focus shifts to how to locate and remove the mis-recognized words from adaptation data.

3.1 Using N-best List to Select Mis-recognized Words

We propose a method through which most of the error words can be located by analyzing the N-best hypotheses. The procedure of the method is as follows:

1. Run N-best decoding on the test utterances. (Note that in unsupervised mode, the enrollment data and the test data are usually the same. In our experiment, we use a total of 332 sentences from 8 speakers as the test data).
2. Pick out the 1-best decoding output as reference transcription **T**.
3. Align and match the remaining N-best outputs with reference **T**. Mark any mismatched words whose mismatch counts exceed a threshold as error words.
4. Weight the marked error word 0 in reference **T**.
5. Run adaptation training using reference **T**.

Compared with the conventional 1st -best adaptation, the proposed method has little impact on computations.

The result of 10-best (threshold is 3) experiment is listed in the last row of Table 3. The proposed method outperforms the conventional 1st -best adaptation:

- A 2% WER reduction was observed.
- The number of rejected sentences reduced 75%.
- 6 out of 8 speakers got improvements and no speaker had degradation.

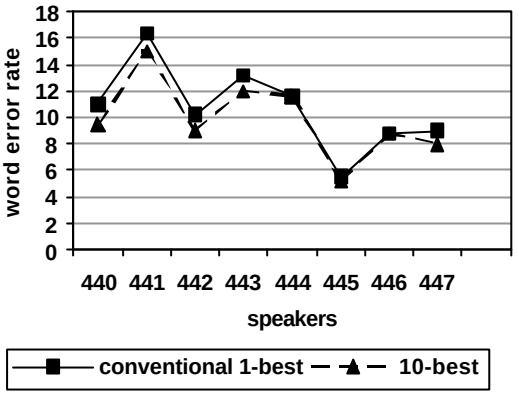

Figure 3. Unsupervised Speaker Adaptations

	# of Rejected Sentences	Average WER
Baseline	5	12.0%
Conventional 1st best	4	10.6%
10-best	1	10.4%

Table 5: Unsupervised Speaker Adaptation

We run a cheating experiment to measure the upper limit of this method. We removed the mis-recognized words from the adaptation data by comparing the best recognition result with the correct reference. Thus we assume we can 100 percent correctly tagging the recognition result as correct or incorrect. The unsupervised adaptation based the above method shows a 9.7% WER. Which implies the room for further improvement is about 0.7% WER reduction.

Another cheating experiment is performing supervised speaker adaptation on the full test set. The result is 7.3% WER on the full test set.

Considering the OOV words cannot be correctly recognized in our situation, it is better to removing their effects when comparing different recognition results. Since there's 2% OOV words in the test data set that causes about 3% WER, we subtracted 3% from all the original WERs. As shown on table 5, the accuracy of classification of correct or incorrect is important for unsupervised adaptation. It's desired to import the confidence measurement technique into current method.

	Average WER	Δ_{WER}
Conventional adaptation	7.6%	
10-best	7.4%	-2.6%
Unsupervised adaptation (cheating)	6.7%	-11.8%
Supervised adaptation (cheating)	4.3%	-43.4%

Table 6: Word Error Rate after removed OOV effects

4. CONCLUSIONS

In this paper, a new adaptation scheme and its application are presented. By weighting each part of the adaptation data differently, the proposed adaptation scheme can use the adaptation data more efficiently and thus improve the system performance. We conclude that in supervised adaptation, the error words should be emphasized, while in unsupervised adaptation the error words should be appropriately identified and discarded. A likelihood weighting function is introduced to the supervised adaptation and an error-word spotting method is proposed for the unsupervised adaptation. However, our error-word spotting method is quite preliminary. Incorporating confidence measures into unsupervised adaptation is being studied.

5. REFERENCES

1. C.L. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs, *Computer Speech and Language*, Vol.9, pp. 171-185, 1995.
2. X.Wu, C. Liu, Y. Yan, D.Kim, S.Cameron, and R.Parr. The 1998 OGI-Fonix Broadcast News Transcription System. *Broadcast News Transcription and Understand Workshop*, 1999.
3. P. Nguyen, Ph. Gelin, J-C. Junqua and J-T. Chien. N-BEST based supervised and unsupervised adaptation for native and non-native speakers in car. *IEEE International Conference on Acoustics, Speech, and Signal Processing*. pp. 173-176, 1999.
4. T. Schaff, T. Kemp. Confidence Measures for Spontaneous Speech Recognition. *Proc. of ICASSP*, pp. 875-878, 1999.
5. M. Siu, H. Gish. Improved Estimation, Evaluation and Applications of Confidence Measures For Speech Recognition. *Proc. of European Conference on Speech Communications and Technology*, pp. 831-834, 1999.
6. M. Weintraub. LVCSR log-likelihood ratio scoring for keyword spotting. *Proc. of ICASSP*, pp. 297-300, 1999.