Iterative Unsupervised Speaker Adaptation for Batch Dictation

Shigeru HOMMA, Jun-ichi TAKAHASHI, and Shigeki SAGAYAMA

NTT Human Interface Laboratories
1-2356 Take, Yokosuka-shi, Kanagawa, 238-03, Japan

ABSTRACT

This paper describes an automatic batch-style dictation paradigm in which the entire dictated speech is fully utilized for speaker adaptation and is recognized using the speaker adaptation results. The key point is that the same speech data is used both for recognition as the target and for speaker adaptation. Two steps, speech recognition and speaker adaptation which uses recognition results as means of supervision, are iterated to maximize the advantage of closed-data speaker adaptation. Recognition errors are reduced by 37% in a practical application of batch-style speech-to-text conversion of recorded dictation of Japanese medical diagnoses compared to speaker-independent recognition. To select only reliable recognition results, a supervision improvement procedure is used by which erroneous recognition results can be eliminated from the supervision. In this procedure, 59-74% of the data are extracted from the tentative recognition results and their reliability is 89-93%. This procedure also reduces recognition errors by 45%.

1. INTRODUCTION

Many dictation systems [1],[2],[6] employing speech recognition have been developed in Europe and America and have been recently gaining popularity in various fields. In the past few years, speaker adaptation has been discussed mainly for use in on-line speech recognition which these systems employ. There are two strategies for adapting the hidden Markov model (HMM) parameters in these systems, batch adaptation and incremental adaptation. These two adaptation strategies have common problems when used for on-line speech recognition: (1) training data does not always provide all possible variations of phonemes of object speech, and (2) the character of object speech is not utilized for adaptive training.

In contrast, off-line speech recognition such as batch-style speech-to-text conversion of a tape-recorded dictation allows another possible mode of speaker adaptation: "off-line, closed-data, unsupervised, batch speaker adaptation" where the entire recorded speech is used for speaker adaptation prior to speech recognition. The advantage of off-line speech recognition is that by fully utilizing the same data for both speaker adaptation and speech recognition leads to significantly better results. Since processing is off-line, fast computational capability for real-time processing is not required. Our approach uses tentative recognition results obtained by recognizing the target speech as means of supervision. Speech recognition and speaker adaptation are alternately performed updating the acoustic models to speaker-dependent models. Using this procedure, effective batch-style speech-to-text conversion of recorded dictation is expected to be achieved only by using the target speech itself similar to closed-data training without any additional data for speaker adaptation.

2. ITERATIVE UNSUPERVISED SPEAKER ADAPTATION

The proposed iterative unsupervised speaker adaptation is illustrated in Figure 1. In this adaptation process, the target data to be dictated are iteratively used for speaker adaptation. The adaptation procedure is summarized below.

1. Speech recognition using initial speaker-independent models (to make tentative supervision corresponding to the speech).
2. Adaptive training similar to closed data training using the recognized phoneme sequences obtained in the previous step.
3. Speech recognition using the latest adapted models (for tentative supervision).
4. If recognition results are different from previous ones, go to 2, otherwise end this procedure.

Due to this iterative closed training using the target data, the initial speaker-independent models are gradually adapted to speaker-specific models.

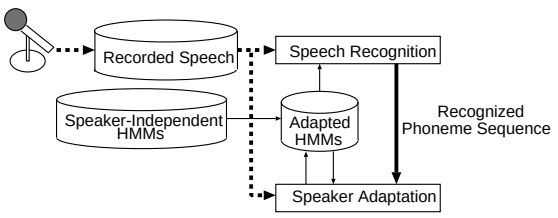

Figure 1: Iterative unsupervised adaptation procedure

2.1. Experimental Setup

This procedure was experimentally evaluated through phoneme-based speaker-independent phrase speech recognition. Maximum likelihood (ML) estimation and maximum a posteriori (MAP) [3],[5] estimation were employed as speaker adaptation algorithms. The initial models were context-dependent speaker-independent models [7]. These models are trained based on the hidden Markov network (HMnet) using a database provided by ATR which consists of 216 phonetically balanced word utterances and 5,240 word utterances by 20 speakers and a database provided by ASJ which consists of 150 phonetically balanced sentences by 64 speakers. This HMnet was constructed using an allophone environment tying technique at triphone-model and state levels. There were approximately 1,700 context-dependent phoneme models in the tied-state configuration with four mixtures in output distributions. The number of states was 450 and the total output distribution counts were 2,280. The feature parameter was a 33-dimension vector consisting of 16 cepstral coefficients, 16 Δ cepstral coefficients, and a Δ log-power. In the experiments, the mean vector was adapted with fixed variances. The grammar of the phrases and LR parser were based on 1,400 reports (approximately 70,000 phrases) concerning head X-ray computerized tomography. The sentence structure was analyzed based on the HMM-LR method [4],[8]. The size of the vocabulary of the dictionary is approximately 3,600 words.

2.2. Experiments and Results

The recognition performance for two female speakers was evaluated through phrase speech recognition experiments. A test set consisting of 30 reports concerning head X-ray computerized tomography (containing 1,300 phrases or 14,000 phonemes) was uttered by each speaker phrase-by-phrase at natural speed. The Recognition performance was evaluated using the percentage of the phrases correctly recognized and is hereinafter called "the phrase conversion accuracy". Phrase conversion accuracy using the latest model was evaluated after each speaker adaptation.

To compare recognition performance, we conducted the following experiments:

- **Batch speaker adaptation**
  (with correct supervision)
  to determine the upper limit of the speaker adaptation
- **Batch speaker adaptation**
  (with correct results of speaker-independent recognition and corresponding speech data)
  to examine recognition performance without erroneous supervision
- **Incremental speaker adaptation**
  (correct supervision is given after each recognition of one sentence or one report)
  to determine the amount of work necessary to correct the recognition error
- **Speaker-independent phrase recognition**
  to determine base recognition performance

Table 1: Comparison phrase conversion accuracy

Adaptation Method	Supervision	Estimation	Conversion Accuracy
Batch	Supervised	ML	94.1%
		MAP	92.6%
	Correct results of speaker-independent recognition	ML	91.1%
		MAP	89.6%
Incremental	every report	MAP	90.7%
	every sentence		90.5%
Iterative Batch	**Unsupervised**	ML	**85.2**%
		MAP	**86.3**%
Speaker Independent	—	—	78.4%

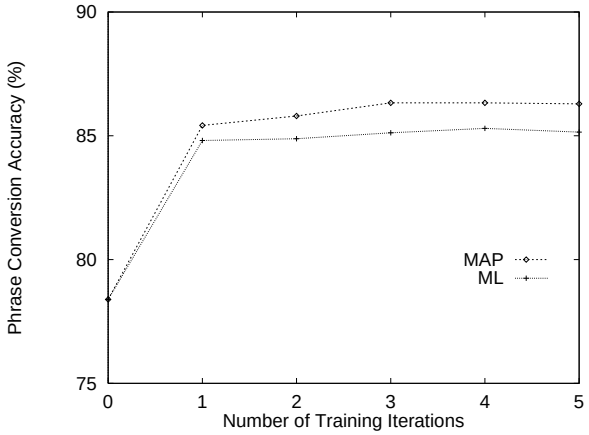

Figure 2: Adaptive learning curves of iterative unsupervised speaker adaptation

Table 1 shows the phrase conversion accuracy of the speaker adaptation strategies. Figure 2 shows the relationship between the phrase conversion accuracy of iterative unsupervised speaker adaptation and the number of iterations. Conversion accuracy rates of iterative unsupervised speaker adaptation with ML and MAP estimation were respectively 85.2% and 86.3% after five iterations. Recognition errors were reduced by 37% compared to speaker-independent recognition.

3. IMPROVED SUPERVISION

In the experiments of iterative unsupervised speaker adaptation, the phrase conversion accuracy is improved only slightly after the second iteration. On the other hand, batch speaker adaptation employing the correct results of speaker-independent recognition and corresponding speech data achieved higher level of phrase conversion accuracy than iterative unsupervised speaker adaptation. The degradation of the adapted speech model was clearly caused by erroneous recognition results. To reduce the affect of erroneous recognition results, there are two possible ways: decreasing the con-

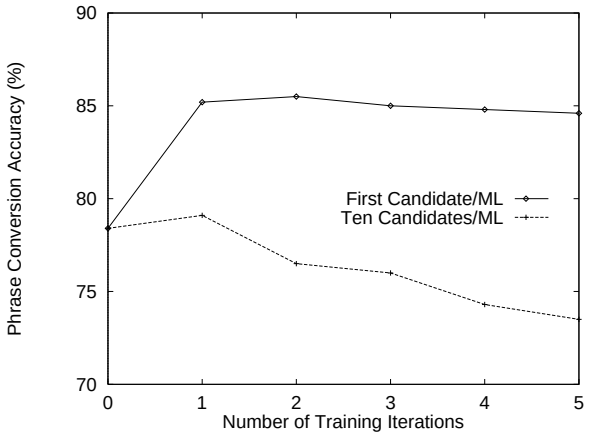

Figure 3: Adaptive learning curves of iterative unsupervised speaker adaptation using posterior probability of candidates in re-estimation of parameters

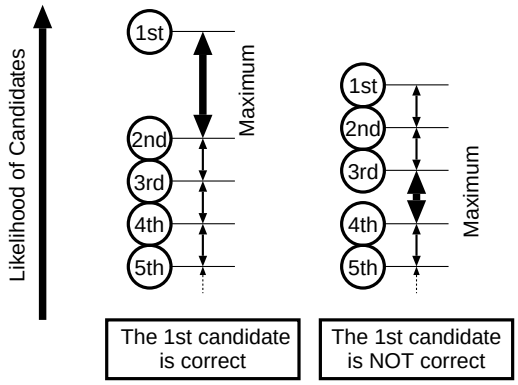

Figure 4: An assumption for supervision improvement procedure

tribution of erroneous recognition results against HMM parameters and eliminating erroneous recognition results from the supervision. As the re-estimation of HMM parameters is based on a probabilistic algorithm, adjusting the contribution of erroneous recognition results from this probabilistic algorithm makes sense. We assumed that when the recognition result is correct, posterior probability of the candidate is high. Using posterior probability among n best candidates to adjust contribution in the the re-estimation of HMM parameters (we used $n = 10$ candidates), we tried to reduce the affect of erroneous recognition results. Two experiments were conducted, one used the first candidates only, and the other used n candidates.

Figure 3 shows the relationship between the phrase conversion accuracy of iterative unsupervised speaker adaptation weighted by posterior probability and the number of iterations. Consequently, eliminating erroneous recognition results from the supervision is the most promising strategy.

3.1. Supervision Improvement Procedure

To select only reliable recognition results, a supervision improvement procedure is used which eliminates erroneous recognition results from the adaptive training information. We assume that when the recognition result is correct, the difference between the likelihood of the first and second candidates tends to be significantly larger than the differences between other adjacent candidates, as illustrated in Figure 4 (candidates have been sorted by their likelihood). Based on this confidence measure, phrase recognition results which conform to this assumption and corresponding speech data are employed for adaptive training.

The supervision improvement procedure is summarized below.

1. Calculate the differences between the likelihood of adjacent candidates ($n = 10$ candidates were used).
2. **Accept** if the difference between the likelihood of the first and second candidates is the largest, otherwise **reject** the recognition result.

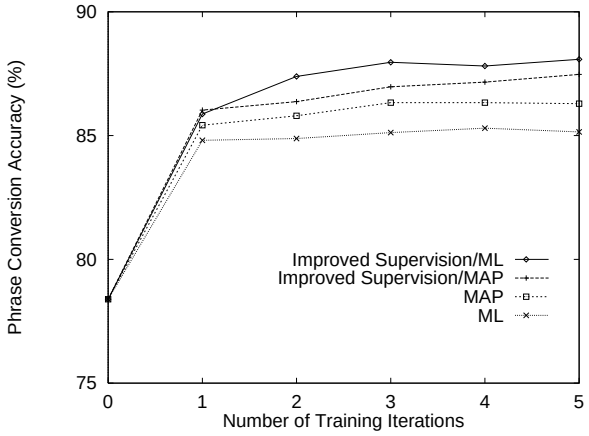

Figure 5: Adaptive learning curves of iterative unsupervised speaker adaptation with improved supervision

Using the supervision improvement procedure the initial speaker-independent models are adapted to the speaker by iterating three steps: (1) recognition of the target speech, (2) selection of the tentative recognition results, and (3) adaptive training of the same recorded dictation speech.

3.2. Experiments and Results

Figure 5 shows the relationship between the phrase conversion accuracy of iterative unsupervised speaker adaptation and the number of iterations. Using the supervision improvement procedure, the conversion accuracy rates were respectively 88.1% and 87.5%, achieving an error reduction rate of 45% compared to speaker-independent recognition. Figure 6 shows the relationship between the utilization of training data and the number of training iterations

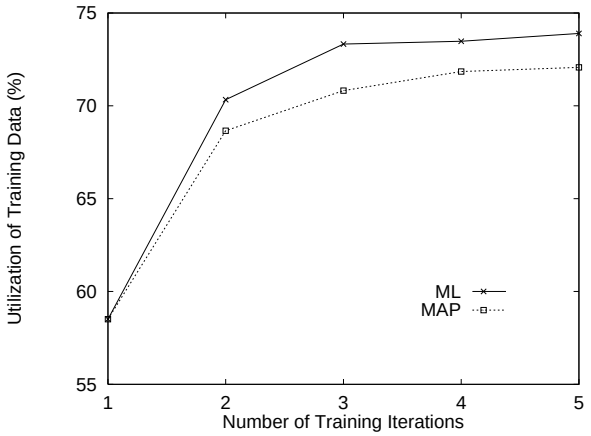

Figure 6: Utilization of training data in iterative unsupervised speaker adaptation with supervision improvement procedure

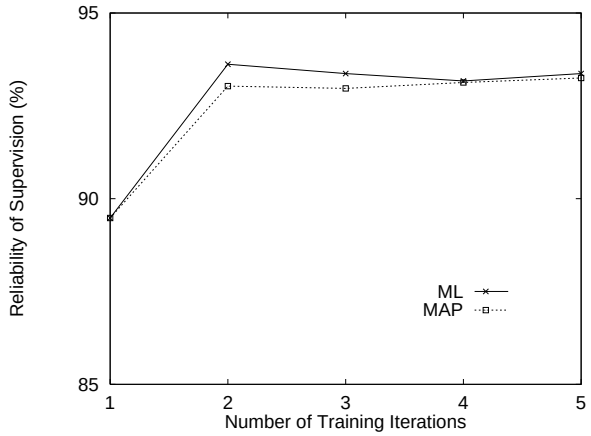

Figure 7: Reliability of supervision selected by supervision improvement procedure in iterative unsupervised speaker adaptation

of the speaker adaptation when the supervision improvement procedure is employed. 59% of the training data was utilized in the first iteration, and in the fifth iteration, 74% was utilized with ML estimation and 72% was utilized with MAP estimation. Figure 7 shows the relationship between the reliability of the supervision (rate of a correct recognition result) and the number of training iterations of the speaker adaptation when the supervision improvement procedure is employed. The reliability of the supervision was 89% in the first iteration and 93% after the second iteration. Clearly, the effective speech-to-text conversion is achieved only by using the target speech itself without any additional data for speaker adaptation. On the other hand, incremental speaker adaptation requires additional information to achieve a slightly higher level of conversion accuracy than our procedure. By increasing the reliability of supervision and utilization of training data, the conversion accuracy of the proposed iterative unsupervised speaker adaptation can be improved to the level of accuracy of batch speaker adaptation which uses the correct results of speaker-independent recognition and corresponding speech data.

4. CONCLUSION

In this paper, we presented an automatic batch-style dictation paradigm in which the entire dictated speech is fully utilized for speaker adaptation and recognized using the speaker adaptation results. In our approach, speaker adaptation is iteratively performed using tentative recognition results obtained by recognition of the target speech as the supervision. This updates the acoustic models to speaker-dependent models. We have also presented a supervision improvement procedure by which correct candidates were determined. When a recognition result is correct, the difference between likelihood of the first and second candidates tends to be significantly larger than the difference between other adjacent candidates. This procedure improves performance by removing incorrect supervision in speaker adaptation. In our performance evaluation, the proposed procedure is very effective in batch-style speech-to-text conversion of recorded dictation tasks.

ACKNOWLEDGEMENTS

We would like to express our gratitude to Dr. Tadayuki Maehara, former head of the Department of Diagnostic Radiology at Kantou Teishin Hospital, for providing samples of head X-ray computerized tomography reports. We would also like to thank Mr. Satoshi Takahashi, who provided programs for HMM training, and Mr. Tomokazu Yamada, who provided programs for recognition.

REFERENCES

[1] A. Averbuch et al. “Experiments with the Tangora 20,000 word speech recognizer”. *Proc. ICASSP’87*, pages 701–704, 1987.

[2] J. Baker. “DRAGONDICTATE-30K: Natural language speech recognition with 30,000 words”. *Proc. Eurospeech’89*, pages 161–163, 1989.

[3] J. L. Gauvain and C. H. Lee. “Bayesian learning of Gaussian mixture densities for hidden Markov models”. *Proc. DARPA Speech and Natural Language Workshop*, pages 272–277, 1991.

[4] K. Kita et al. “HMM continuous speech recognition using predictive LR parsing”. *ICASSP’89*, pages 703–706, 1989.

[5] C. H. Lee, C. H. Lin, and B. H. Juang. “A study on speaker adaptation of the parameters of continuous density hidden Markov models”. *IEEE Trans. ASSP*, 39(4):806–814, 1991.

[6] V. Steinbiss et al. “The Philips research system for large-vocabulary continuous-speech recognition”. *Eurospeech’93*, pages 2125–2128, 1993.

[7] S. Takahashi and S. Sagayama. “Four-level tied-structure for efficient representation of acoustic modeling”. *ICASSP’95*, pages 520–523, 1995.

[8] M. Tomita. *Efficient parsing for natural language; a fast algorithm for practical system*. Kluwer Academic Publishers, 1986.