IMPROVED ESTIMATION OF SUPERVISION IN UNSUPERVISED SPEAKER ADAPTATION

Shigeru Homma, Kiyoaki Aikawa, and Shigeki Sagayama

NTT Human Interface Laboratories
1–1 Hikarinooka, Yokosuka-Shi, Kanagawa 239, Japan
{honma,aik,saga}@nttspch.hil.ntt.co.jp

ABSTRACT

Unsupervised speaker adaptation plays an important role in "batch dictation," the aim of which is to automatically transcribe large amounts of recorded dictation using speech recognition. In the case of unsupervised speaker adaptation which uses recognition results of target speech as the means of supervision, erroneous recognition results degrade the quality of the adapted acoustic models. This paper presents a new supervision selection method. By using this method, correction of the first candidate is judged based on the likelihood ratio of the first and the second candidates. This method eliminates erroneous recognition results and corresponding speech data from the adaptive training data. We implemented this method in the iterative unsupervised speaker adaptation procedure. It is shown that the recognition errors are drastically reduced by 50% in a practical application of batch-style speech-to-text conversion of recorded dictation of Japanese medical diagnoses compared with speaker-independent recognition.

1. INTRODUCTION

Dictation systems [1],[2],[3] employing speech recognition have been developed in Europe and America and have recently been 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 object speech itself 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 speaker adaptation strategy: "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 significantly better results can be obtained. Since processing is performed off-line, fast computational capability for real-time processing is not required. Our approach [4] used 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 toward speaker-dependent models. 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 a higher level of phrase conversion accuracy than did iterative unsupervised speaker adaptation. Erroneous recognition results degrade the quality of the adapted acoustic models.

There are two possible ways to reduce the effect of erroneous recognition results: decreasing the contribution of erroneous recognition results when re-estimating HMM parameters and eliminating erroneous recognition results from the supervision. Since the re-estimation of HMM parameters is based on a probabilistic algorithm, adjusting the contribution of erroneous recognition results from this probabilistic algorithm is a practical method. We assumed that when the recognition results are correct, the posterior probability of the candidate is high. Using posterior probability among n best candidates to adjust the contribution in the re-estimation of HMM parameters, we tried to reduce the effect of erroneous recognition results. However, this idea did not work well. Consequently, eliminating erroneous recognition results from the supervision was the most promising strategy. To select only reliable recognition results, we assumed 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. Based on this confidence measure, phrase recognition results which conform to this assumption and corresponding speech data are employed for adaptive training. By eliminating erroneous recognition results from the supervision, effective speech-to-text conversion was achieved. However, the reliability of supervision and utilization of training data were not high enough so far.

2. IMPROVEMENT OF SUPERVISION

The most important issue for unsupervised speaker adaptation is training efficiency, which relies on collecting as much training data as possible and reducing the harmful influence of erroneous speech labels mistakenly selected from among recognition results. Using the efficiency rating is an ideal way to satisfy both the high reliability of speech labels and the high rate utilization of training data when selecting correct results from among speaker-independent recognition results for the use of unsupervised speaker adaptation. In general, these conditions are properly compromised. We determined, through preliminary experiments, the condition for maximizing the efficiency of adaptive training.

2.1. Reliability of Supervision and Utilization of Training Data

In the preliminary experiments, the effects of the reliability of supervision and utilization of training data in speaker adaptation training were examined by artificially changing their values. After adaptive training through pseudo-supervision, phrase recognition experiments were conducted using adapted models. The recognition performance was

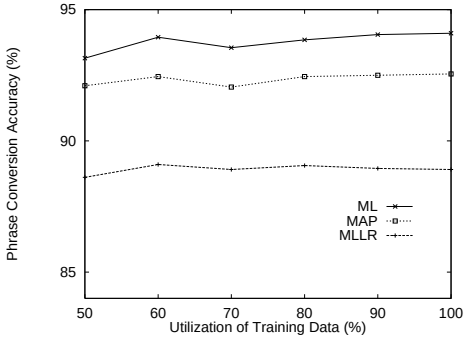

Figure 1. Relationship between utilization of training data and phrase conversion accuracy

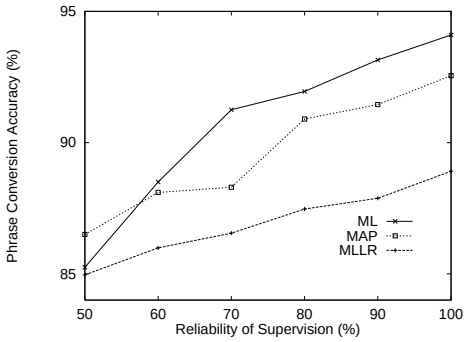

Figure 2. Relationship between reliability of supervision and phrase conversion accuracy

expressed using the percentage of the phrases correctly recognized and is called "the phrase conversion accuracy". The Maximum Likelihood (ML) estimation, Maximum *A Posteriori* (MAP)[5],[6] estimation, and Maximum Likelihood Linear Regression (MLLR)[7],[8] estimation were employed as speaker adaptation training methods. The number of regression classes of MLLR was 6. These regression classes were pre-determined prior to adaptation and fixed based on the phoneme. Figure 1 shows the relationship between the utilization of training data and phrase conversion accuracy. The respective inclinations of the graphs of ML, MAP, and MLLR were 0.015, 0.008, and 0.0034 by the least squares. Figure 2 shows the relationship between the reliability of the supervision and the phrase conversion accuracy. The respective inclinations of the graphs of ML, MAP, and MLLR were 0.17, 0.12, and 0.075 by the least squares. Errors in supervision are from ten to twenty times more detrimental than the decrease in training data when comparing the inclinations of the graphs of the same employed estimation method.

2.2. Judgment of Correction of Recognition Results using the Likelihood Ratio

The likelihood ratio was used to verify decoded utterances in order to account for incorrectly decoded vocabulary words and utterances corresponding to words or sounds that are not included in a prespecified lexicon, and the effectiveness was previously reported (e.g.,[9],[10]). When the first candidate of a speech recognition result is correct, the likelihood of the second or the other candidates tend to be smaller than the that of the first candidate, if the character of the object speaker is not completely different from those of the speakers who are represented in the training set of

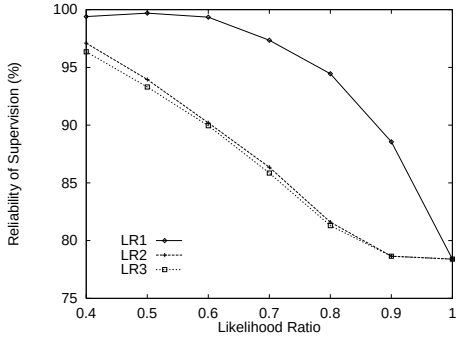

Figure 3. Relationship between likelihood ratio and reliability of supervision

speaker-independent models. Based on this idea, correction of the first candidate of the speech recognition result can be judged by the value of the likelihood ratio of the first and the other candidates. If the likelihood ratio (LR) is smaller than the decision threshold, the first candidate of recognition result is judged correct. The likelihood ratios are defined as:

$$LR1 = \frac{L_2}{L_1} \quad (1)$$

$$LR2 = \exp\left(\frac{1}{N-1}\sum_{n=2}^{N}\log\left(\frac{L_n}{L_1}\right)\right) \quad (2)$$

$$LR3 = \left(\frac{1}{N-1}\sum_{n=2}^{N}\left(\frac{L_n}{L_1}\right)^{-\kappa}\right)^{-\frac{1}{\kappa}} \quad (3)$$

where L_n is the likelihood of the nth candidates of the speech recognition result.

If the decision threshold of the likelihood ratio can be appropriately determined, high reliability of supervision and comparatively high utilization of training data are attained at once for the selected training data in the unsupervised speaker adaptation training procedure. The effect of the decision threshold on the reliability of supervision and the utilization of training data were examined using speaker-independent recognition results in the preliminary experiments. Figure 3 shows the relationship between the likelihood ratio and the reliability of supervision. The reliability of training data is the percentage of correct recognition results among used data. Figure 4 shows the relationship between the likelihood ratio and utilization of training data. As mentioned in 2.1., the cost of errors in the supervision were ten or more times more detrimental than the decrease in training data, therefore $LR1$ was employed for the selection of training data. To determine the appropriate threshold, we employed "the Bayesian decision rule for minimum risk" (e.g., [11]) which minimizes the expected cost (EC)

$$EC = c_{12}\pi_1 P_{12} + c_{21}\pi_2 P_{21} \quad (4)$$

where c_{ij} is cost of determining $X \in \omega_j$ when $\in \omega_i$, π_i is the a priori probability of $X \in \omega_i$, and P_{ij} is the probability of error in determining $X \in \omega_j$ when $X \in \omega_i$ respectively. The threshold which minimizes the EC is determined by the intersection of two likelihood ratio distributions weighted by cost and a priori probability.

The likelihood ratio distribution when the first candidate of the recognition result is correct and when it is in error were examined using speaker-independent recognition results in the preliminary experiments. Figure 5 shows distributions of the weighted likelihood ratio. As mentioned

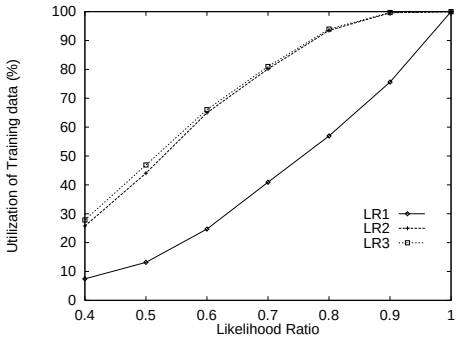

Figure 4. Relationship between likelihood ratio and utilization of training data

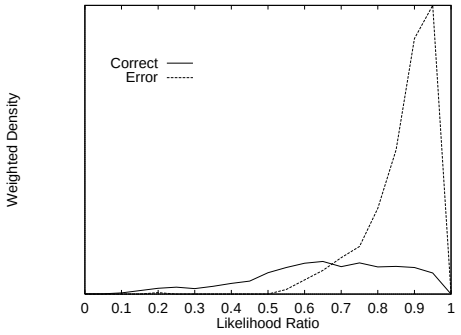

Figure 5. Distributions of weighted likelihood ratio when the first candidate of recognition result is correct and when it is in error

in 2.1., the cost of errors in the supervision were set to ten times the cost of the decrease of training data. As a result, the threshold was determined to be 0.7 using the intersection of two likelihood ratio distributions. When LR is smaller than 0.7, the possibility of the first candidate of recognition result being correct will be 97% and the utilization of the training data will be 41%, as a result of the preliminary experiments in 2.2..

2.3. New Supervision Selection Method

To select only reliable recognition results, a new supervision selection method is used which eliminates erroneous recognition results from the adaptive training information. The new supervision selection method is summarized below.

1. Calculate the LR of the first and the second candidates of the speech recognition result.
2. **Accept** the recognition result and the corresponding training data for adaptive training, if the LR does not exceed the decision threshold (0.7), otherwise **reject** the recognition result.

3. EXPERIMENTS

The presented supervision selection method was implemented in an iterative unsupervised speaker adaptation procedure[4] and was evaluated using phoneme based speaker-independent phrase recognition. Using the new supervision selection method the initial speaker-independent models were adapted to the target speaker by iterating the following three steps as shown in figure 6.

1. Recognition of the whole target speech using the latest model.

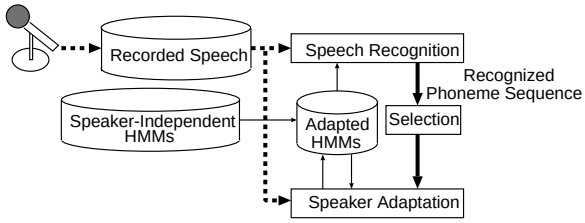

Figure 6. Iterative unsupervised speaker adaptation procedure

2. Selection of the tentative recognition results using new supervision selection method.
3. Adaptive training of the same recorded dictation speech using selected recognition results and corresponding speech data.

3.1. Experimental Setup

The initial models were context-dependent speaker-independent models [12]. 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 the Acoustical Society of Japan (ASJ) which consists of 150 phonetically balanced sentences by 64 speakers. This HMnet was constructed using an allophone environment tying technique at the triphone-model and state levels. This HMnet was equivalent to 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 task was a medical diagnosis concerning X-ray CT scanning of a human head. To obtain grammatical information for an LR parser [13],[14], training texts of 1,400 reports including a substantial number of phrases, 70,000, were used. The dictionary included approximately 3,600 words. The speech of 30 CT scanning reports dictated by each of two female speakers were used as the target data including approximately 1,300 phrases.

3.2. Results

Figure 7 shows the relationship between the phrase conversion accuracy of iterative unsupervised speaker adaptation using the new supervision selection method and the number of iterations. Phrase conversion accuracy values of 88.1%, 88.9%, and 89.3% were achieved for the first, second, and third iterations, respectively, when ML adaptation was employed. When using MAP adaptation, they were 88.3%, 87.7%, and 88.3 % respectively. When using MLLR adaptation, the accuracy values were 85.6%, 85.7%, and 85.5% respectively. The phrase conversion accuracy of the speaker-independent recognition was 78.4%. The maximum error reduction rate of the conventional iterative unsupervised speaker adaptation without using the new supervision selection method was 37% compared to speaker-independent recognition. According to these results, the error reduction rate increased by 13% from 37% to 50% when ML adaptation was employed due to the new supervision selection method. The reliability of the supervision was approximately 97% while that of the conventional iterative unsupervised speaker adaptation was the conversion accuracy itself. Figure 8 shows the relationship between the utilization of training data and the number of training iterations

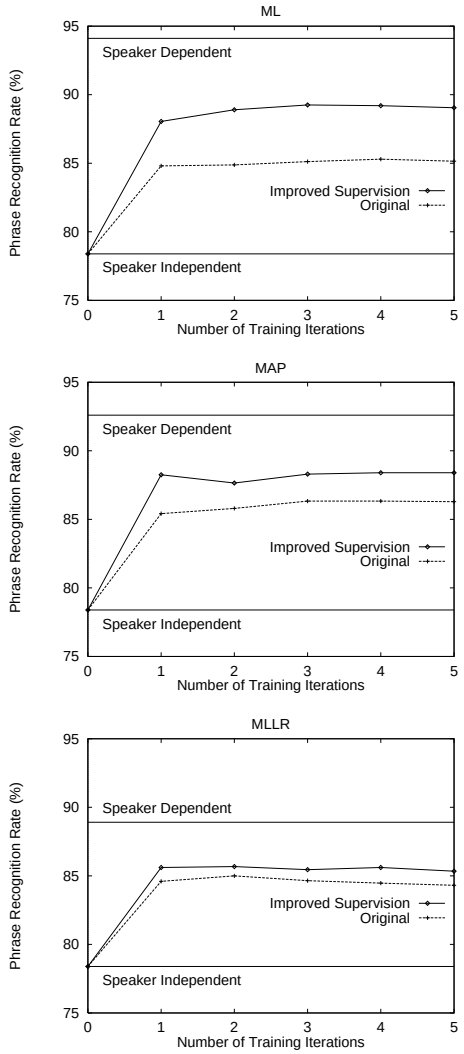

Figure 7. Adaptation training curves in phrase recognition

of the speaker adaptation when the new supervision selection method is employed. 41% of the training data was utilized in the first iteration. By increasing the reliability of supervision, the conversion accuracy of the iterative unsupervised speaker adaptation was improved though the utilization of training data decreased.

4. CONCLUSION

In this paper, we presented a new supervision selection method. Using this method, correction of the first candidate was judged based on the likelihood ratio of the first and the second candidates. This method eliminated erroneous recognition results and corresponding speech data from the adaptive training data. We implemented this method in the iterative unsupervised speaker adaptation procedure and evaluated it through phrase recognition experiments using medical diagnoses data. As a result, the error reduction rate increased by 13% from 37% to 50% compared with the conventional iterative unsupervised speaker adaptation without new supervision selection method. This result shows that the presented method improved the efficiency of unsupervised speaker adaptation training.

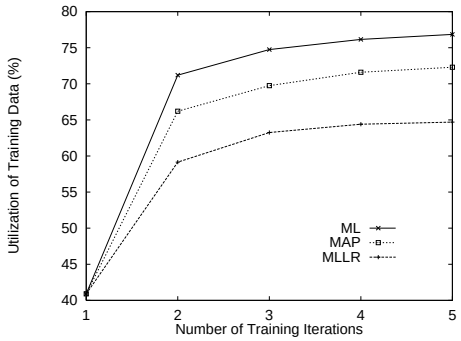

Figure 8. Utilization of training data in iterative unsupervised speaker adaptation with new supervision selection method

5. ACKNOWLEDGMENTS

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*, pp. 701–704, 1987.

[2] J. Baker, "DRAGONDICTATE-30K: Natural language speech recognition with 30,000 words,"*Proc. Eurospeech'89*, pp. 161–163, 1989.

[3] V. Steinbiss et al., "The Philips research system for large-vocabulary continuous-speech recognition," *Eurospeech'93*, pp. 2125–2128, 1993.

[4] S. Homma, J. Takahashi, and S. Sagayama, "Iterative unsupervised speaker adaptation for batch dictation," *Proc. ICSLP'96*, pp. 1141–1144, 1996.

[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*, Vol. 39, No. 4, pp. 806–814, 1991.

[6] J. L. Gauvain and C. H. Lee, "Bayesian learning of Gaussian mixture densities for hidden Markov models," *Proc. DARPA Speech and Natural Language Workshop*, pp. 272–277, Arden House, 1991.

[7] C. J. Leggetter and P. C. Woodland, "Speaker Adaptation of Continuous Density HMMs using Multivariate Linear Regression," *Proc. ICSLP'94*, pp. 451–454, 1994.

[8] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," Computer Speech and Language, Vol. 9, No. 2, pp. 171–185, 1995.

[9] E. Lleida and R. C. Rose, "Efficient decoding and training procedures for utterance verification in continuous speech recognition," *Proc. ICASSP'96*, pp. 507–510, 1996.

[10] R. A. Sukkar, et al., "Utterance verification of keyword strings using word-based minimum verification error (WB-MVE) training," *Proc. ICASSP'96*, pp. 518–521, 1996.

[11] K. Fukunaga, *Introduction to statistical pattern recognition*, Academic Press, 1972.

[12] S. Takahashi and S. Sagayama, "Four-level tied-structure for efficient representation of acoustic modeling," *ICASSP'95*, pp. 520–523, 1995.

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

[14] K. Kita, et al., "HMM continuous speech recognition using predictive LR parsing," *Proc. ICASSP'89*, pp. 703–706, 1989.