SMOOTHED N-BEST-BASED SPEAKER ADAPTATION FOR SPEECH RECOGNITION

Tomoko Matsui, Tatsuo Matsuoka, and Sadaoki Furui

NTT Human Interface Laboratories
3-9-11, Midori-cho, Musashino-shi, Tokyo, Japan

ABSTRACT

Smoothed estimation and utterance verification are introduced into the N-best-based speaker adaptation method. That method is effective even for speakers whose decodings using speaker-independent (SI) models are error-prone, that is, for speakers for whom adaptation techniques are truly needed. The smoothed estimation improves the performance for such speakers, and the utterance verification reduces the required amount of calculation. Performance evaluation using connected-digit (four-digit strings) recognition experiments performed over actual telephone lines showed a reduction of 36.4% in the error rates for speakers whose decodings using SI models are error-prone. To try and find an effective model-transformation for speaker adaptation, we discuss replacing mixture-mean bias estimation by the widely used mixture-mean linear-regression-matrix estimation.

1. INTRODUCTION

In continuous mixture-density hidden Markov model (HMM)-based speech-recognition systems, the performance of speaker-independent (SI) phoneme HMMs for some speakers is often poor. Techniques that adapt the parameters of SI phoneme HMMs to each speaker and thus improve the performance are therefore important. These techniques are usually classified as supervised or unsupervised, in which either training utterances with or without the transcriptions are used, respectively. They can also be classified either off-line or on-line.

Instantaneous adaptation is unsupervised and on-line: the recognition utterances are used to estimate the adaptation transformation. It is especially useful in applications where there is only a very brief interaction between the speaker and the system [1][2]. This technique must work using only a small amount of data, such as a few words or a single sentence.

In general, unsupervised adaptation techniques use a recognized word sequence, W^*, obtained using SI phoneme HMMs. Parameter set θ of the SI phoneme HMMs is adapted to each speaker by finding the control parameter set η in model transformation function $G_\eta(\theta)$ that maps θ to the speaker-adapted parameter set according to the following equation, which is based on, for instance, maximum a posteriori (MAP) estimation [3].

$$\tilde{\eta} = \arg\max_{\eta} f(X|W^*, \eta, \theta) g(\theta), \tag{1}$$

where X is an observation sample, $f(\cdot)$ is the likelihood function, and $g(\cdot)$ is the a priori density function. Speaker-adapted parameter set $\tilde{\theta}$ is calculated using $\tilde{\eta}$:

$$\tilde{\theta} = G_{\tilde{\eta}}(\theta).$$

However, recognition is error-prone for some speakers, and the adaptation usually does not work for those speakers.

Our proposed N-best-based instantaneous speaker-adaptation method is effective even for error-prone speakers [4]. This method finds $\tilde{\eta}$ and $\tilde{W}$ such that

$$(\tilde{\eta}, \tilde{W}) = \arg\max_{(\eta, W)} f(X|W, \eta, \theta) g(\theta) P(W), \tag{2}$$

where $P(W)$ is the a priori probability of the word sequence W. (In the connected-digit recognition task used in this paper, $P(W)$ is assumed to be constant for all word sequences and is thus ignored.)

Because it is too costly to attempt speaker adaptation for all possible word sequences, to reduce the search space without losing the correct sequence, the N-best paradigm of multiple-pass search strategies [5] is used to calculate likely sequences. In this method, hierarchical adaptation with several estimation iterations is used to constrain η in order to avoid estimating η at an unexpected local maximum. In each iteration, the adapted phoneme models for the sequence that has the highest likelihood value are selected and used in the next iteration. However, the correct word sequence does not always show the highest likelihood value in the earlier estimation iterations where adaptation is insufficient. In particular, when the correct word sequence of the input speech is recognized as one of the lower best using SI phoneme HMMs, the speech is often decoded incorrectly even when adaptation is used.

We have thus added a smoothed estimation technique in which not only the sequence with the highest likelihood value after adaptation but also the other sequences are taken into account. In this technique, a confidence measure for each sequence is defined according to its likelihood value after adaptation; value of each model parameter is calculated as the weighted sum of the values of the model parameters for all N-best sequences using the confidence

$$b_n^i = \frac{\sum_{\{j,k\,|\,\phi(j,k)=i\}} [\tau_{jk}(\mu_{jk} - m_{jk}) + \sum_{t=1}^{T} c_{jkt}^n (x_t - m_{jk})] r_{jk}}{\sum_{\{j,k\,|\,\phi(j,k)=i\}} (\tau_{jk} + \sum_{t=1}^{T} c_{jkt}^n) r_{jk}} \tag{3}$$

measures. We have also added an utterance verification technique that reduces the required amount of calculation. In our original method, to cope with the insufficient amount of data, mixture-mean bias estimation based on MAP was used. To try and find an effective model-transformation for speaker adaptation, we discuss replacing mixture-mean bias estimation by the widely used mixture-mean linear-regression-matrix estimation based on maximum likelihood linear-regression (MLLR) [6].

2. SMOOTHED N-BEST ADAPTATION

The proposed method consists of three steps:

1. Multiple word-sequence hypotheses of the input speech $\{W_1, W_2, \ldots, W_N\}$ are obtained using SI phoneme HMMs by using N-best decoding.
2. The parameters of the phoneme models are adapted for each decoding by finding η_n:

$$\eta_n = \arg\max_{\eta} f(X|W_n, \eta, \theta) g(\theta),$$

$$\theta_n \leftarrow G_{\eta_n}(\theta).$$

3. A confidence measure $\mathcal{C}_n$ is defined for instance:

$$\mathcal{C}_n = \frac{1}{\exp(\alpha[\tilde{\mathcal{L}_{max}} - \tilde{\mathcal{L}_n}])},$$

where $\tilde{\mathcal{L}_{max}}$ is the highest log-likelihood value after adaptation, $\tilde{\mathcal{L}_n}$ is the log-likelihood value for W_n after adaptation and α is an experimental parameter. After $\mathcal{C}_n$ is calculated for each decoding, the parameters of the phoneme models are calculated:

$$\tilde{\theta} \leftarrow \frac{\sum_{n=1}^{N} \mathcal{C}_n \theta_n}{\sum_{n=1}^{N} \mathcal{C}_n}.$$

The difference between our original method [4] and this enhanced version is in Step 3. Originally, in Step 3, the decoding providing the maximum-likelihood value was selected for the speech, and the speaker-adapted phoneme models for that decoding were used.

Steps 2 and 3 in both methods are iterated until the adaptation in Step 2 becomes sufficiently precise by using the hierarchical codebook adaptation algorithm [7][8]. This algorithm was developed for speaker adaptation in vector-quantization-based systems: the reference codebook elements are clustered hierarchically by increasing the number of clusters, and adaptation to the speaker is performed hierarchically from the global individuality characteristics down to the local ones. In practice, the mixture-mean bias model, in which the biases are shared by the distributions in the same cluster, is used for model transformation function $G_\eta(\cdot)$, and the number of biases or matrices (i.e., the number of clusters) increases as the number of estimation iterations increases.

In Sections 2.1 and 2.2, we explain how the mixture means are clustered and the mixture-mean biases are estimated.

2.1. Hierarchical clustering

In our method, a binary tree-structure is created from the input speech. The number of estimation iterations corresponds to the depth of the tree, and the number of leaves on each level corresponds to the number of clusters. The mixture-density distributions are classified into 2^{M-1} clusters based on the distances between the centroids and the mean vectors of the distributions, where M is the number of estimation iterations.

2.2. Bias estimation

For each sequence W_n in Step 2, the mixture-mean bias set $\{b_n^1, b_n^2, \ldots, b_n^I\}$ for clusters 1 to I is approximated while $f(X|W_n, \{b_n^i\}, \theta) g(\theta)$ is locally maximized using MAP estimation, where b_n^i is given by

$$G_{b_n^{\phi(j,k)}}(m_{jk}) = m_{jk} + b_n^{\phi(j,k)}. \tag{4}$$

Here, m_{jk} is the mean vector of the mixture component k in state j, and $\phi(\cdot)$ is a membership function indicating the cluster to which the mixture component in the state belongs. The expectation-maximization reestimation formula is shown in Eq. (3), where μ_{jk} and τ_{jk} are the a priori density parameters, r_{jk} is the precision vector, and c_{jkt} is the probability of observation vector x_t generated by the HMM at time t being in state j with mixture component k. To maintain continuity between the clusters, the bias for each mixture mean of all phoneme HMMs is calculated as the weighted sum of the biases $\{b_n^1, b_n^2, \ldots, b_n^I\}$ based on the distances between the centroids and the mixture-mean vector.

3. EXPERIMENTAL EVALUATION

3.1. Conditions

The database we used for creating SI (four-mixture Gaussian) HMMs consisted of Japanese digit-strings (one, two, and four digits) spoken by 177 male speakers (24,194 strings in total). The data was collected by NTT over actual telephone lines in a metropolitan area. In our experiments, each digit was represented using sub-word HMMs (head, body, and tail models [9][10]) or using whole-word HMMs, depending on the context. We used 100 sub-word HMMs (43 head, 13 body, and 54 tail models) and 13 whole-word HMMs to represent the digits 0 to 9.

The data used for adaptation and recognition testing consisted of four-digit strings spoken by 50 male speakers randomly selected from a different database consisting of data collected by NTT Data over actual telephone lines in seven different areas [11]. Six different strings were used per speaker. The ten best hypotheses were decoded using SI digit models, and the percentage of correct strings

Method	Difficult spkr		Easy spkr	Avg.
Baseline	59.3	-	93.9	87.7
1-best	63.0	[10.1]	94.7	89.0
10-best	70.4	[27.3]	92.7	88.7
Smoothed 10-best	74.1	[36.4]	92.7	89.3

Table 1. String recognition rate (%) for several adaptation methods ([ ]: error reduction rate (%)).

included within these ten best decodings was 97.3%. Sixteen mixture-mean biases were estimated for each string by using hierarchical clustering with five estimation iterations ($M_{max} = 5$). The 50 speakers were classified into two sets: one set consisted of 9 "difficult" speakers, each of whom had two or more strings recognized incorrectly using the SI HMMs; the other set consisted of the remaining 41 "easy" speakers, each of whom had one or zero strings recognized incorrectly.

The 12th-order cepstral and delta-cepstral coefficients were calculated. Linear predictive coding analysis was used with a frame period of 8 ms and a frame length of 32 ms. Cepstral mean subtraction was performed for each utterance.

3.2. Results

Table1 lists the string recognition rates and the error reduction rates compared with the baseline performance. In the "1-best" method, the recognized sequence is used for adaptation in the conventional way. The "10-best" is our original method [4], and the "Smoothed 10-best" is our enhanced version. The Smoothed 10-best method was the most effective for difficult speakers whose decodings using SI models were error-prone (36.4% error reduction).

4. INCORPORATION OF UTTERANCE VERIFICATION

We also added an utterance verification technique to reduce the required amount of calculation for our method, in which the adaptation for one utterance must be applied N times for each iteration. For strings recognized correctly using SI models (set A) and for strings recognized incorrectly using SI models but correctly through adaptation (set B), we examined the log-likelihood ratios between the likelihood values for strings recognized as best and second best using SI models (Figure1). We found that the log-likelihood ratios for set B were small and that 90.5% of the strings included in set A had higher log-likelihood ratios than the highest log-likelihood ratio for set B. For those strings it is not necessary to use N-best decoding; that is, adaptation can be done by simply using best decoding without any degradation in performance. Therefore, if utterance verification, in which the threshold shown by the dashed line in Figure1 (B) is used in order to judge whether the N-best decoding is necessary or not, is performed, the calculation amount for our method can be greatly reduced.

Table2 lists the calculation reduction rates in real time and and the string recognition rates when using the above utterance verification (UV) with a posteriori threshold. Utterance verification not only reduced the calculation time, but also improved the performance for both our original and enhanced methods. The average recognition rate for the Smoothed 10-best method increased from 89.3% to 90.0% (the error-reduction rate compared with the baseline performance increased from 13.0% to 18.7%).

In N-best-based adaptation, strings recognized correctly using SI models may become misrecognized through adaptation, probably because adaptation by constrained mixture-mean bias-estimation is not sufficient for some utterances, so the correct word sequences do not show the highest likelihood values. In our experiments described in Section 3, the number of utterances so affected was seven for the 10-best method and six for the Smoothed 10-best method. With utterance verification, this number was reduced to four in both cases.

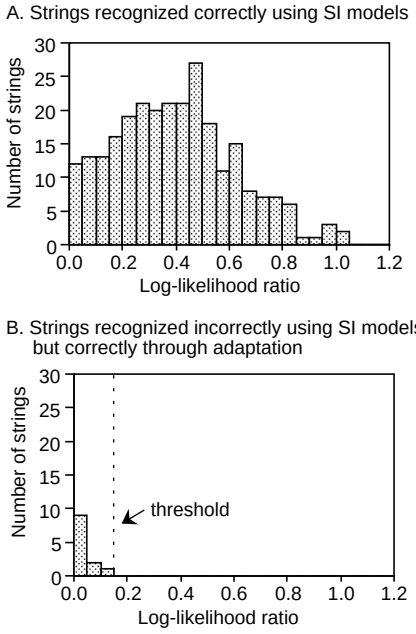

Figure 1. Histograms of log-likelihood ratios.

Method	Calc. red. rate	Avg. rec. rate	
		without UV	with UV
10-best	77.9	88.7	89.7
Smoothed 10-best	80.1	89.3	90.0

Table 2. Calculation reduction rate (%) in real time and improved recognition rate (%) with utterance verification (UV).

5. BIAS VS. LINEAR-REGRESSION MATRIX ESTIMATION

To find an effective model-transformation function for speaker adaptation, we also conducted N-best adaptation experiments using linear-regression (LR) model for the

Method	Difficult spkr	Easy spkr	Avg.
LR 10-best	70.4 [27.3]	93.9	89.7
LR 10-best with UV	72.2 [31.7]	93.9	90.0

Table 3. String recognition rate (%) for linear-regression matrix estimation ([ ]: error reduction rate (%)).

model-transformation function. For each sequence W_n in Step 2 in Section 2, we approximated the $(p+1) \times p$-LR-matrix set $\{[a_n^1, b_n^1], [a_n^2, b_n^2], \ldots, [a_n^I, b_n^I]\}$ (p : dimension of observation vector, a_n^i : $p \times p$-matrix, b_n^i : p [dimensional]-vector) while locally maximizing $f(X|W_n, \{[a_n^i, b_n^i]\}, \theta)$ by using MLLR [6], where $[a_n^i, b_n^i]$ is given by

$$G_{[a_n^{\phi(j,k)}, b_n^{\phi(j,k)}]}(m_{jk}) = a_n^{\phi(j,k)} m_{jk} + b_n^{\phi(j,k)}. \qquad (5)$$

Table3 lists the string recognition rates for LR-matrix estimation without and with utterance verification when a global LR diagonal matrix was estimated in one estimation iteration. Although the LR 10-best method with UV performed as well as the Smoothed 10-best method with UV using bias-estimation (Table2) on average for all speakers, we believe that the main need is to improve the performance for difficult speakers. As shown in Table1 , the Smoothed 10-best method produced the biggest increase in performance for difficult speakers. The LR 10-best method is a promising approach because fewer parameters need to be estimated (a $p \times p$-diagonal-matrix plus a p-vector, that is, 2 p-components) than in the Smoothed 10-best method (16 p-vectors). However, so far we have been unable to improve the performance of the LR 10-best method by increasing the number of LR matrices or by applying smoothed estimation in one iteration. This is probably because of the incontinuity between clusters after adaptation. We are now studying methods to maintain continuity and to estimate multiple LR matrices hierarchically and smoothly over several iterations.

6. CONCLUSION

We have presented an N-best-based instantaneous speaker adaptation method with smoothed estimation for continuous mixture-density HMM-based speech-recognition systems. Connected-digit (four-digit strings) recognition experiments performed over actual telephone lines showed that this method, which can work with only a small amount of data, is especially effective for difficult speakers whose decodings using SI models are error-prone. We also showed how the use of utterance verification reduces the required amount of calculation and reduces the number of strings that become misrecognized through adaptation. Comparison of the performance between mixture-mean bias estimation and LR-matrix estimation showed that both methods were equally effective on average, but smoothed bias-estimation was more effective for difficult speakers. With the former approach, the error-reduction rate for difficult speakers was 36.4%, and the average for all speakers was 18.7%.

7. ACKNOWLEDGMENTS

We are grateful to the NTT Data Communications Systems Corporation for allowing us to use their database for our experiments. We thank to Ken Hatano and Naoki Hashimoto of the Tokyo Institute of Technology for their valuable assistance with our experiments. We also thank the members of the Furui Research Laboratory of the NTT Human Interface Laboratories for their valuable and stimulating discussions.

REFERENCES

[1] G. Zavaliagkos, R. Schwartz and J. Makhoul, *Batch, incremental and instantaneous adaptation techniques for speech recognition*, Proc. ICASSP, pp. I-676-679, 1995.

[2] A. Sankar, Leonardo Neumeyer and M. Weintraub, *An experimental study of acoustic adaptation algorithms*, Proc. ICASSP, pp. II-713-716, 1996.

[3] J.L. Gauvain and C.-H. Lee, *Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains*, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 2, pp. 291-298, 1994.

[4] T. Matsui and S. Furui, *N-best based instantaneous speaker adaptation method for speech recognition*, Proc. ICSLP, pp. III-973-976, 1996.

[5] R. Schwartz, L. Nguyen and J. Makhoul *Automatic speech and speaker recognition: Chapter 18 Multiple-pass search strategies*, edited by C.-H. Lee et al., Kluwer Academic Publishers, pp. 429-456, 1995.

[6] 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, pp. 171-185, 1995.

[7] Y. Shiraki and M. Honda, *Speaker adaptation algorithm based on piecewise moving adaptive segment quantization method*, Proc. ICASSP, pp. II-657-660, 1990.

[8] S. Furui, *Unsupervised speaker adaptation based on hierarchical spectral clustering*, IEEE Trans. on ASSP, Vol. 37, No. 12, pp. 1923-1930, 1989.

[9] R. Pieraccini, C.-H. Lee, E.Giachin and L.R. Rabiner, *Implementation aspects of large vocabulary recognition based on intraword and interword phonetic units*, Proc. DARPA Speech and Natural Language Workshop, pp. 311-318, 1990.

[10] T. Matsuoka, N. Uemoto, T. Matsui and S. Furui, *Elaborate acoustic modeling for Japanese connected digit recognition*, Proc. IEEE Automatic Speech Recognition Workshop, Snowbird, pp. 169-170, 1995.

[11] M. Morishima, T. Isobe and K. Murakami, *Telephone speech database and the experiment using CDHMM*, Proc. Acoustical Society of Japan, Fall Meeting, 2-8-8, 1994.