On-line Incremental Adaptation Based on Reinforcement Learning for Robust Speech Recognition

Masafumi Nishida, Yoshitaka Mamiya, Yasuo Horiuchi, and Akira Ichikawa

Graduate School of Science and Technology
Chiba University, Japan
{nishida, hory, ichikawa}@faculty.chiba-u.jp

Abstract

We propose an incremental unsupervised adaptation method based on reinforcement learning in order to achieve robust speech recognition in various noisy environments. Reinforcement learning is a training method based on rewards that represents correctness of outputs instead of supervised data. The training progresses gradually based on rewards given. Our method is able to perform environmental adaptation without priori knowledge about such things as speakers and noises in noisy environments. We conducted speech recognition experiments using a connected digit recognition database. We demonstrate that our method has higher recognition performance than the conventional adaptation method.

1. Introduction

We are aiming for robust speech recognition in real world applications such as dialogue systems. In these applications, the system needs to deal with a variation of both speakers and noises. The purpose of this study is to create speech recognition that performs adaptation without prior knowledge about such things as speakers and noises and that improves recognition accuracy by using the system repeatedly.

Many approaches have been proposed to handle the problem of mismatch between training and application conditions. These can be classified into two categories: signal processing techniques and compensation schemes. In the former, Missing feature theory [1] identifies non-reliable time-frequency regions of speech, and the marginalization method is used for missing components. Multi-band speech recognition [2] is a paradigm in which each frequency region is treated as a distinct source of information, and the streams are combined after each is processed independently. However, the signal processing approach is not able to adapt to a variation of speakers.

In the latter method, initial canonical models (generally clean speech models) are transformed to represent a new environment. The stochastic matching method [3] offers some chance of reducing the mismatch by mapping distorted features to an estimate of the original features or mapping original models to transformed models. The Parallel Model Combination (PMC) method [4] can derive noisy speech HMMs by combining a clean speech HMM and a noise HMM based on the SNR. However, the method needs noise data beforehand and combining the PMC with other mismatch compensation techniques such as cepstrum mean normalization is difficult. Maximum A Posteriori (MAP) estimation [5] is a model compensation method based on an amount of adaptation data and priori density. The Maximum Likelihood Linear Regression (MLLR) method [6] can deal with additive noise and convolitional distortion simultaneously. The adaptation performance by these methods decreases if transcription accuracy is poor in the unsupervised adaptation because it performs the adaptation according to a fixed weight for all transcription results. However, the MAP estimation is more suitable for incremental learning than the MLLR method.

We propose an incremental unsupervised adaptation method based on reinforcement learning that adapts to the variation of both speakers and noises. Reinforcement learning [7] is a training method based on rewards that represents correctness of outputs instead of supervised data. The training progresses gradually based on rewards given. The proposed method adapts a variation of environments by controlling the weight of priori density in the MAP according to the rewards.

We conducted speech recognition experiments using AURORA-2 Japanese corpus [8] (connected digit recognition database) to show the effectiveness of the proposed method.

2. Reinforcement Learning

Reinforcement learning [7] is learning how to map situations to actions so as to maximize a numerical reward signal. The learner is not told which actions to take, as in most forms of machine learning, but instead must discover which actions yield the most reward by trying them. The actions may affect not only the immediate reward,
but also the next situation and, through that, all subsequent rewards.

Reinforcement learning is different from supervised learning. Supervised learning is learning from examples provided by some knowledgeable external supervisor. This is an important kind of learning, but alone, it is not adequate for learning from interaction. In interactive problems, it is often impractical to obtain examples of desired behavior that are both correct and representative of all the situations in which the agent has to act.

One of the challenges that arises in reinforcement learning and not in other kinds of learning is the tradeoff between exploration and exploitation. To obtain a lot of rewards, an agent must perform actions that it has tried in the past and found to be effective in producing rewards. But to discover such actions, it has to try actions that it has not selected before. The agent has to exploit what it already knows to obtain the reward, but also has to explore in order to make better actions in the future.

We consider that it is possible to perform an unsupervised environmental adaptation without priori knowledge about such things as speakers and noises by applying reinforcement learning to the recognition of noisy speech.

3. Acoustic Adaptation Based on Reinforcement Learning

We describe a method that adapts acoustic models to speakers and environments using reinforcement learning. Acoustic adaptation based on reinforcement learning is achieved based on an incremental MAP adaptation.

The MAP is a method to estimate a model parameter \( \theta \) when posteriori probability \( P(\theta|x) \) is maximum using observation data \( x \). The posteriori probability is defined by Bayesian theory.

\[
P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}
\]

(1)

The parameter \( \theta \) is unrelated to Priori density \( P(x) \) when it is estimated using the observation data. Thus, in a hypothesis where the parameter \( \theta \) is a probabilistic variable based on the priori density, parameter \( \bar{\theta} \) is estimated as follows.

\[
\bar{\theta} = \arg \max_\theta P(\theta|x) = \arg \max_\theta P(x|\theta)P(\theta)
\]

(2)

In a hypothesis where the priori density is based on normal density \( N(\mu_0, \sigma_0) \), the mean vector of mixtures of HMM is estimated by the MAP,

\[
\bar{\mu} = \frac{1}{\alpha + T} \{ \alpha \mu_0 + \sum_{i=1}^{T} x_i \}
\]

(3)

where \( x_i \) is a feature vector of adaptation data, \( T \) is the number of frames of adaptation data, and \( \alpha \) is an updating coefficient. When the mean vector is adapted by Eq. (3), the adapted mean vector \( \bar{\mu} \) approaches the direction of adaptation data rather than the mean vector \( \mu_0 \) before updating. The parameter \( \alpha \) is able to control the rate of priori density, and the weight of past adaptation decreases by incremental adaptation.

Next, we describe the adaptation method of acoustic models based on reinforcement learning. In this paper, input and output data are digit sequences because we perform a connected digit recognition under noisy environments. A reward that represents correctness of outputs for the input data is needed to adapt HMMs by reinforcement learning. Thus, we extract digit sequences when likelihood is maximum for each frame and digit sequences obtained by the Viterbi algorithm. We investigate whether both digit sequences are the same or different for each frame by comparing these digit sequences. Then, the reward is determined based on the compared results. In the adaptation process, the reward is given to the updating coefficient \( \alpha \) in Eq. (3).

A flow diagram of our method is shown in Fig. 1. The detailed procedure is as follows.

1. The likelihood of HMMs is computed for the input data, and digit sequences with a maximum likelihood are obtained for each frame (a in Fig. 1).
2. After likelihoods are computed for all frames, digit sequences with maximum likelihood for a whole utterance are obtained by the Viterbi algorithm (b in Fig. 1).
3. The digit sequences obtained by the Viterbi algorithm are compared with the digit sequences for each frame (c in Fig. 1).
4. Acoustic models are adapted based on the MAP according to the reward given by the compared results. (d in Fig. 1)

A comparison method of the digit sequences is shown in Fig. 2. In the figure, the upper row is the digit sequences with the maximum likelihood for each frame,
and the lower row is the digit sequences obtained by the Viterbi algorithm. The digit sequences for each frame are a direct output from HMMs, and the digit sequences by the Viterbi algorithm reflect the frame order. Thus, we assume that the confidence of the recognition results is high in the speech section when the digit sequences determined by the Viterbi algorithm are the same as the digit sequences for each frame. Therefore, we assume that the confidence of the recognition results is high in the speech section when the digit sequences determined by the Viterbi algorithm are the same as the digit sequences for each frame. Thus, we assume that the confidence of the recognition results is high in the speech section when the digit sequences determined by the Viterbi algorithm are the same as the digit sequences for each frame. Therefore, the MAP adaptation is performed using a small updating coefficient $\alpha$ in the speech sections (I), (III), and (IV) in Fig. 2 because both digit sequences are the same. In the speech sections (II) and (V) in Fig. 2, the MAP adaptation is performed using a large updating coefficient $\alpha$ because both digit sequences are different.

It is possible to carry out an unsupervised adaptation incrementally by our method with variation in speakers and environments.

4. Experiments

4.1. Experimental Conditions

We conducted speech recognition experiments using AURORA-2 Japanese corpus [8]. The corpus is a connected digit recognition database. Each utterance contains 1 to 7 digits.

The speech data was sampled at 16 kHz and segmented into 25 ms frames every 10 ms with 15 ms overlap. The acoustic features consist of 39 components of 12 MFCCs, energy, and their first and second order derivatives. The cepstrum mean normalization was applied to each utterance. Each digit was modeled by a 16-state HMM, and each state has a mixture of 20 diagonal Gaussians. The initial acoustic models were trained using clean speech data. The training data consists of 8440 utterances from 110 speakers.

To evaluate our method, we used speech data where 8 kinds of noises (Subway, Babble, Car, Exhibition, Restaurant, Street, Airport, and Station) were added to test data with three SNRs: 10, 15, and 20 dB. The test data contains 500 utterances from 50 speakers. The speaker is different for every 10 utterances.

4.2. Experimental Results

The recognition accuracy for each SNR is shown in Table 1. “Baseline” denotes the result when the baseline model that was trained using clean speech data was used without adaptation. “MAP” denotes the result when an unsupervised adaptation by the conventional MAP was performed incrementally. The updating coefficient $\alpha$ was set to 30 in a preliminary experiment. “Reinforcement learning” denotes the result when the unsupervised adaptation based on reinforcement learning was performed incrementally. In the preliminary experiment, the updating coefficient $\alpha$ was set to 30 when the digit sequences determined by the Viterbi algorithm were the same as the digit sequences for each frame. The coefficient $\alpha$ was set to 70 when the both digit sequences were different.

<table>
<thead>
<tr>
<th>SNR</th>
<th>Baseline</th>
<th>MAP</th>
<th>Reinforcement learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 dB</td>
<td>89.4</td>
<td>96.8</td>
<td>97.0</td>
</tr>
<tr>
<td>15 dB</td>
<td>65.8</td>
<td>90.1</td>
<td>91.7</td>
</tr>
<tr>
<td>10 dB</td>
<td>43.0</td>
<td>48.3</td>
<td>61.3</td>
</tr>
</tbody>
</table>

The conventional MAP adaptation method outperformed the method using the baseline model. However, improvement in recognition accuracy was small for the case of 10 dB. This is because the MAP method performs the adaptation according to a fixed weight for all recognition results, although recognition errors increase when the SNR decreases.

The proposed method achieved a recognition accuracy of 97.0% for the case of 20 dB, 91.7% for 15 dB, and 61.3% for 10 dB. It outperformed the method using the baseline model and the conventional MAP adaptation method. It achieved the best performance of all the SNRs. The method obtained a significantly higher improvement rate than the MAP adaptation when the SNR was 10 dB. This demonstrates that it is effective to adapt using an optimal reward based on a confidence of recognition results.

The recognition accuracy for each noise is shown in Fig. 3. Figure 3 denotes the average recognition accuracy of the SNRs for each noise. The improvement was observed by the proposed method for almost all noises.

The recognition accuracy for every 20 utterances is shown in Fig. 4. Figure 4 denotes the average recognition accuracy of all noises with 10 dB. In our method, the recognition accuracy was improved gradually, and the variation of the recognition accuracy was smaller than the other methods. This demonstrates that incremental unsupervised adaptation based on reinforcement learning is effective in adapting the variation of both speakers and noises.
5. Conclusion

We proposed an on-line incremental adaptation method based on reinforcement learning for robust speech recognition. The method was able to adapt to variation of both speakers and noises without priori knowledge about such things as speakers and noises. We conducted speech recognition experiments using a connected digit recognition database. We demonstrated that our method achieves higher recognition performance than conventional methods such as MAP adaptation for all noises.

We will study the estimation of an optimal reward value and a system using a combination of our method and other approaches such as noise reduction methods.

6. References


