REINFORCEMENT LEARNING OF SPEECH RECOGNITION SYSTEM BASED ON POLICY GRADIENT AND HYPOTHESIS SELECTION

Taku Kato, Takahiro Shinozaki, Tokyo Institute of Technology, Japan

Overview

- Background
  - Today's automatic speech recognition (ASR) systems heavily rely on supervised training using large amounts of task-matched training data
  - The cost of transcribing speech data is repeatedly required to support new languages and new tasks
  - A system would become more self-sufficient and useful if it possessed the ability to learn from very light feedback from the users
- Our contribution
  - Formulate a general reinforcement learning framework for ASR systems based on the policy gradient method
  - Propose a hypothesis selection method following the reinforcement learning framework, where the feedback is given by user selection of hypotheses selection

Related work

- User based correction of recognition errors in cloud environment
  - PodCastle [Ogata et al., Interspeech, 2007]
  - Laborious effort is required

Policy Gradient (PG) Method

- Assumptions
  - We have a policy function f with a set of parameters $\boldsymbol{\theta}$
  - Input : A state or observation $\boldsymbol{s}$
  - Output : A probability distribution $P_f(a|\boldsymbol{s})$ of an action a
  - Reward $r_s(a)$ is given for the action
- Goal
  - Maximize the expected reward $\mathbb{E}[r_s(a)]$ with respect to $\boldsymbol{\theta}$
- Gradient ascent based solution

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}[r_s(a)|\boldsymbol{\theta}] = \mathbb{E}[r_s(a)\nabla_{\boldsymbol{\theta}} \log P_f(a|\boldsymbol{s})]$$

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \varepsilon r_s(a)\nabla_{\boldsymbol{\theta}} \log P_f(a|\boldsymbol{s}) \quad \varepsilon : \text{The learning rate}$$

- General form of REINFORCE algorithm [Williams, 1992]

$$(r-b)\frac{\partial \log g(i)}{\partial \boldsymbol{\theta}}$$

b : Reinforcement baseline
$g(i)$: Neural network based policy function
$\boldsymbol{\theta}$: Parameters of the neural network

Formulation of PG for statistical ASR systems

- Input $\boldsymbol{s}$: A feature sequence of an utterance
- Action : A probability distribution of a word sequence l of recognition hypothesis
- Policy function : The whole statistical ASR system

- The probability distribution

$$P(\boldsymbol{l}|\boldsymbol{s}) = \frac{P_{AM}(\boldsymbol{s}|\boldsymbol{l})P_{LM}(\boldsymbol{l})}{P(\boldsymbol{s})} \propto \underbrace{\boxed{P_{AM}(\boldsymbol{s}|\boldsymbol{l})}}_{\text{update}} P_{LM}(\boldsymbol{l})$$

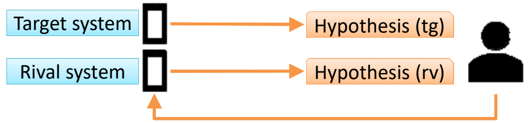
Acoustic model — Language model

- The gradient

$$r_s(a)\nabla_{\boldsymbol{\theta}} \log P_f(a|\boldsymbol{s}) = r_s(\boldsymbol{l})\nabla_{\boldsymbol{\theta}} \log P_{AM}(\boldsymbol{l}_t|\boldsymbol{s}_t)$$

Design of user feedback

- Accuracy-based feedback
  - Calculating word accuracy is difficult and time consuming for the user
- Selection-based feedback (Proposed method)
  - Two recognition systems present hypotheses to the user
  - The user selects the better hypothesis among them



$$\text{Reward } r = \begin{cases} 1 : \text{Select hypothesis (tg)} \\ 0 : \text{Select hypothesis (rv)} \end{cases}$$

Implementation with Approximation

- Hypothesis generation : Sampling from posterior distribution → Viterbi decoding
- Rival system
  → Use the n-th $(1 \le n)$ best hypothesis of the same system as the rival hypothesis
  - Hypothesis(tg) : The Candidate 1 hypothesis $l^{(1)}$
  - Hypothesis(rv) : The Candidate 2 hypothesis $l^{(2)}$
- Parameter update : Utterance based update → Large batch based update

Weighted gradient

$$\boxed{(1+\alpha)\left(r - \frac{\alpha}{1+\alpha}\right)\frac{\partial \log P_{AM}(\boldsymbol{l}_t^{(1)}|\boldsymbol{s}_t)}{\partial \boldsymbol{\theta}}} \quad \text{Candidate 1}$$

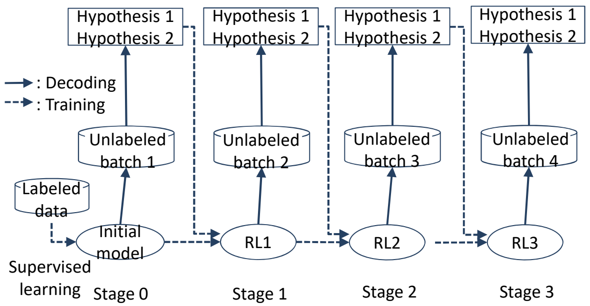
$$+\boxed{(1+\alpha)\left((1-r) - \frac{\alpha}{1+\alpha}\right)\frac{\partial \log P_{AM}(\boldsymbol{l}_t^{(2)}|\boldsymbol{s}_t)}{\partial \boldsymbol{\theta}}} \quad \text{Candidate 2}$$

$\alpha \ (0 \le \alpha \le 1)$: A scalar constant

$$\begin{cases} \dfrac{\partial \log P_{AM}(\boldsymbol{l}_t^{(1)}|\boldsymbol{s}_t)}{\partial \boldsymbol{\theta}} - \alpha \dfrac{\partial \log P_{AM}(\boldsymbol{l}_t^{(2)}|\boldsymbol{s}_t)}{\partial \boldsymbol{\theta}} & (r=1) \\ \dfrac{\partial \log P_{AM}(\boldsymbol{l}_t^{(2)}|\boldsymbol{s}_t)}{\partial \boldsymbol{\theta}} - \alpha \dfrac{\partial \log P_{AM}(\boldsymbol{l}_t^{(1)}|\boldsymbol{s}_t)}{\partial \boldsymbol{\theta}} & (r=0) \end{cases}$$

Increase the difference of the likelihood between the selected hypothesis and the other hypothesis

Learning Process

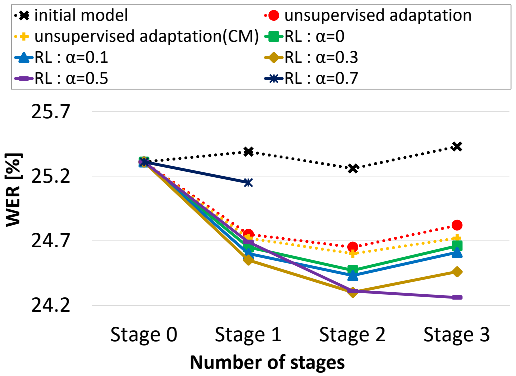


Experimental Conditions

Database	Corpus of Spontaneous Japanese (CSJ)
Training set (labeled)	10 hours
Training set (unlabeled)	50 + 50 + 50 + 50 hours
Evaluation set	2 hours
Vocabulary size	72k words
Initial learning rate	0.004, 0.002, 0.001 and 0.0005
Decoder	Kaldi toolkit
Candidate 2 hypotheses	10-best results
Baseline (unsupervised adaptation)	Confidence measure (CM) based hypothesis selection (Keeps 75% of the hypotheses)

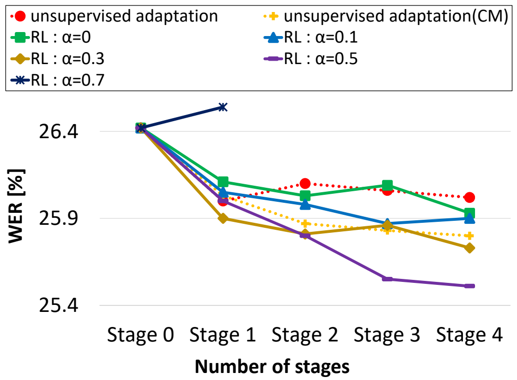
Results (without Hypotheses Selection Error)

Number of stages and WERs of the large batch data



Cf. When supervised training was performed, the WER at stage 3 was 19.3%

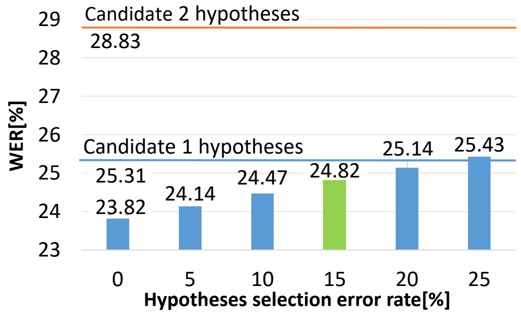
Number of stages and WERs of the evaluation set



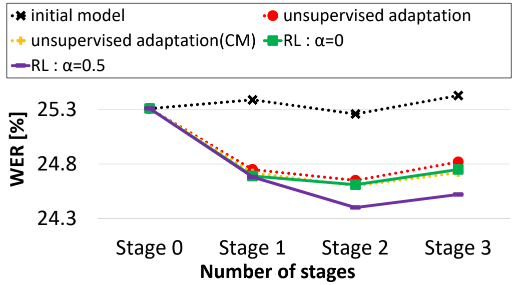
Cf. When supervised training was performed, the WER at stage 4 was 20.6%

Results (with Hypotheses Selection Error)

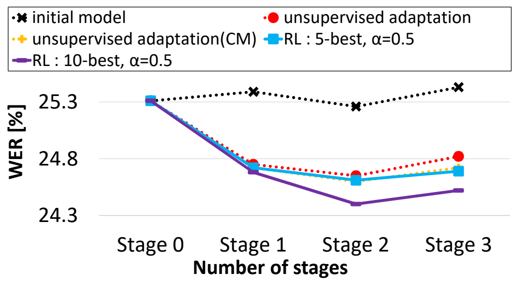
Hypotheses selection error rate and WER of the selected hypotheses



WER of the large batches when 15% hypotheses selection error exist



N-best order of the 2nd hypothesis and WER. 15% selection error rate is simulated



Conclusions

- Formulated a policy gradient-based reinforcement learning framework for ASR systems, and proposed a hypothesis selecting-based reinforcement learning method
- The proposed method reduced WER compared to the unsupervised adaptations
- Future work : Improving the stability to over-training and the learning efficiency for the user feedback