로그 우도 차이의 p-norm에 기반한 은닉 마르코프 파라미터 추정 알고리듬

*윤성락, 유창동
한국과학기술원 전자전산학과 전기 및 전자 공학 전공
e-mail : *yunsungrack@kaist.ac.kr, cdyoo@ee.kaist.ac.kr*

The p-Norm of Log-likelihood Difference Estimation Algorithm for Hidden Markov Models

***Sungrack Yun, Chang D. Yoo**
Department of Electrical Engineering
Korea Advanced Institute of Science and Technology

Abstract

This paper proposes a discriminative training algorithm for estimating hidden Markov model (HMM) parameters. The proposed algorithm estimates the parameters by minimizing the p-norm of log-likelihood difference (PLD) between the utterance probability given the correct transcription and the most competitive transcription.

I. 서론

The discriminative training such as minimum classification error (MCE) and maximum mutual information (MMI) has better performance than the maximum likelihood (ML) for estimating the HMM parameters. The proposed algorithm is based on the approximated MMI and generalizes it from 1-norm of log-likelihood difference form to the p-norm of log-likelihood difference form.

II. 본론

The approximated MMI objective function [1] is given by

$$F(\theta) = \sum_{u=1}^{U} \left[\log P_\theta(\mathbf{O}_u|w_u) - \lambda \log P_\theta(\mathbf{O}_u|v_u) \right]$$

where U is the number of training utterance and $\log P_\theta(\mathbf{O}_u|w_u)$ is the likelihood of u-th utterance $\mathbf{O}_u$ given the correct transcription w_u. v_u is the 1-best recognition transcription of $\mathbf{O}_u$, and λ controls the discrimination rate. Since v_u is the 1-best recognition transcription,

$$\log P_\theta(\mathbf{O}_u|v_u) - \log P_\theta(\mathbf{O}_u|w_u) > 0.$$

In realistic speech recognition task, the probability $P_\theta(\mathbf{O}_u|v_u)$ and $P_\theta(\mathbf{O}_u|w_u)$ becomes very low value, and their logarithm are negative. Thus, the following inequality is satisfied for $0 < \lambda < 1$,

$$\lambda \log P_\theta(\mathbf{O}_u|v_u) - \log P_\theta(\mathbf{O}_u|w_u) > 0.$$

With the above inequality, the objective function of the p-norm of log-likelihood difference (PLD) is obtained by multiplying the weight

$$W_u^{p-1} = \left[\lambda \log P_\theta(\mathbf{O}_u|v_u) - \log P_\theta(\mathbf{O}_u|w_u) \right]^{p-1}$$

to the negative of the approximated MMI objective function. The objective is minimizing the p-norm of log-likelihood difference. The PLD places more weight on the larger log-likelihood difference term so that the parameter update equations are highly dependent on it. The larger log-likelihood difference means that it has more error than the smaller log-likelihood difference.

Ⅲ. 구현

The implementation of the proposed algorithm is based on the method in [1] and generalize it for this algorithm. Since the expectation-maximization and Baum-Welch algorithm are used, the objective function increases monotonically for every iteration.

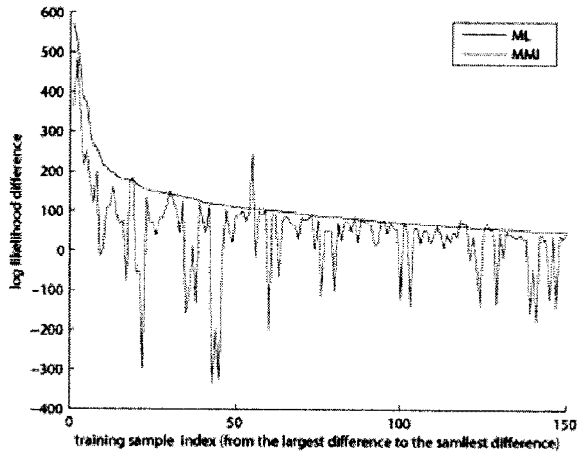

Figure 1. The log-likelihood difference reduction when approximated MMI is used

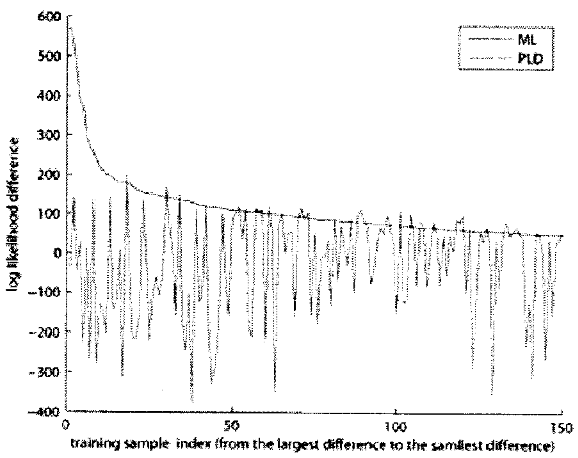

Figure 2. The log-likelihood difference reduction when PLD is used

The experiments were performed to evaluate the PLD: connected digit recognition using the TIDIGIT database. In the experiments, the PLD improved the recognition rate over the approximated MMI and the maximum likelihood (ML). The recognition result is shown in Table 1. Zero insertion penalty was used. In figure 1 and 2, the log-likelihood difference reductions are shown. The PLD reduces the log-likelihood difference more than the MMI for the larger difference term.

# of Gaussian Mixtures	ML	approximated MMI	PLD
1	90.17	91.92	93.05
2	94.21	95.33	96.18
4	95.42	96.60	97.17
8	96.60	97.12	97.68
16	97.36	97.89	98.11
32	97.46	98.00	98.28

Table 1. Recognition rate (%) of ML, approximated MMI and PLD

Ⅳ. 결론 및 향후 연구 방향

Based on the approximated MMI objective function, the proposed algorithm generalizes it from 1-norm of log-likelihood difference form to the p-norm of log-likelihood difference form. With controlling the weighting function, the recognition rate can increase monotonically. The recognition rate decreases for some weighting function because the parameter update equations are only dependent on few dominant terms with large likelihood difference.

참고문헌

[1] A. B. Yishai and D. Burshtein, "A discriminative training algorithm for hidden Markov models," IEEE Trans. Speech and Audio Processing, vol. 12, no. 3, pp. 204 - 217, 2004.