Vol.2010-SLP-82 No.13
2010/7/24

An Evaluation of Discriminative Training for Hidden Markov Models in a Real-Environment Speech-Oriented Guidance System

Denis Babani,[†1] Tomoki Toda,[†1]
Hiroshi Saruwatari[†1] and Kiyohiro Shikano[†1]

This paper presents experimental evaluations of discriminative training of the acoustic model in a real-environment speech-oriented guidance system. Recently, discriminative training techniques have made significant progress in automatic speech recognition (ASR) based on the hidden Markov model (HMM). Their effectiveness has been confirmed in well-known ASR data sets. It is also worthwhile to investigate its effectiveness in more challenging speech data sets. In this paper, we evaluate the effectiveness of discriminative training in speech data recorded in a real environment speech-oriented guidance system, "Takemaru-kun", which has been installed in a public place in November 2002 and has recorded input speech data since then. The recorded speech data include very spontaneous speech of various speakers such as children, adults, and elderly people. Maximum Mutual Information (MMI) training is implemented for building HMMs using these speech data. First we investigate the performance of discriminative training by changing initial conditions in the acoustic model structure and lattice generation. Then, we optimize several training parameters such as the acoustic scale factor and the I-smoothing parameter. Our results show that MMI training yields around 2% absolute word accuracy improvement compared with ML training in the speech data recorded in the "Takemaru-kun" system.

1. Introduction

Automatic speech recognition (ASR) has been an active research field for more than four decades. Accuracy of ASR has been significantly improved according to the implementation of statistical approaches and an increase of the amount of available spoken audio and text data. Consequently, it has made possible the implementation of a new type of human-machine interaction, speech-based natural user interface (NUI). One of the typical applications using speech-based NUI is a spoken dialog system (SDS). SDS will be widely used for various services such as an informational service, an educational service, or even an entertainment service.

We have developed a real environment speech-oriented guidance system, "Takemaru-kun"[1)], based on SDS techniques consisting of large-vocabulary continuous speech recognition (LVCSR), inquiry classification, and text-to-speech (TTS) components. This system has been installed in a public place in November 2002. Since then, really spontaneous utterances of various age groups from infants to elderly people have been collected in real environment for more than seven years. The use of the system will be interesting and from time to time fun. However, its performance is still far from desired.

There have been proposed many techniques that have made significant progress in LVCSR based on the hidden Markov model (HMM) such as model adaptation[2),3)], noise compensation[4),5)], precision modeling[6),7)], and discriminative training[8)–11)]. Their effectiveness has been already confirmed in well-known ASR data sets. It is worthwhile to apply these state-of-the-art HMM-based acoustic modeling techniques to real environment applications and evaluate their effectiveness in real case scenarios.

In this work we investigate the benefits of using Maximum Mutual Information (MMI) training[8),11)], one of the well-known discriminative training methods, in "Takemaru-kun" using speech data recorded by the same system. We evaluate the performance of MMI training by changing various conditions. The experimental results demonstrate that MMI training also yields significant improvements of word accuracy in "Takemaru-kun" system.

This paper is organized as follows. In section 2, we introduce "Takemaru-kun" system and speech data collected in the system. In section 3, MMI discriminative training is described. In section 4, we conduct experimental evaluations of MMI in the "Takemaru-kun" speech data and discuss various implementation issues such as initial conditions and training parameters. Finally, we summarize this paper in section 5.

†1 Nara Institute of Science and Technology

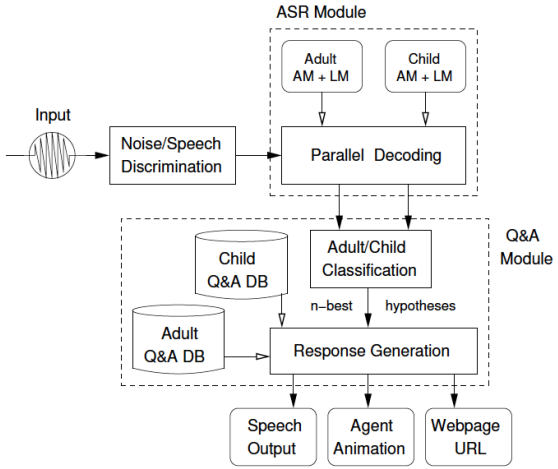

Fig. 1 Process of "Takemaru-kun" system.

2. Speech-Oriented Guidance System "Takemaru-kun"

2.1 Overview of the System

"Takemaru-kun" is a speech-guidance system installed in North community center of Ikoma city, Nara prefecture, Japan. The purpose of this system is to handle queries related to the agent, general information and about surrounding area. "Takemaru-kun" has been up and running for more than 7 years, allowing us to build a rich corpus of spontaneous Japanese utterances.

The process of "Takemaru-kun" system is shown in fig. 1. For each user's request the system goes four consecutive states. As the first state, the system labels the input signal (voice signal) as noise or speech. Then it goes to the next state which is speech recognition if the input signal is labeled as speech. Recognition is done in parallel using two different acoustic models. One acoustic model is trained using adults data and the other using children data. In this state N-best hypothesis are generated for each acoustic model (adults and children) and are served as input of question and answering (Q&A) module. First, this module selects the best input, generated using adults or children acoustic model. Based on the best input, Q&A module generates an appropriate response for the input request. In the last state the system generates sound output, webpage display, and agent animation based on the generated response and presents them to the user.

2.2 Takemaru-kun Data

The number of speech utterances gathered from Takemaru continues to grow faster each year, meanwhile transcribed utterances take more time and effort to be prepared. Until now we have only the first two years completely transcribed by humans. These utterances have been labeled as speech, noise or partially speech and have been subjectively grouped into five groups related to age of the speaker (i.e. preschool, lower grade school children, higher grade school children, adults and elderly persons). Group classification and the number of transcribed utterances is shown in fig. 2.

Utterances recorded by Takemaru are usually short in length. Being in a real environment, speech data usually is not clean, it contains microphone noise, background noise, and even speech overlapping between multiple speakers. For this reason, this corpus is adequate for evaluating MMI performance in real case scenarios.

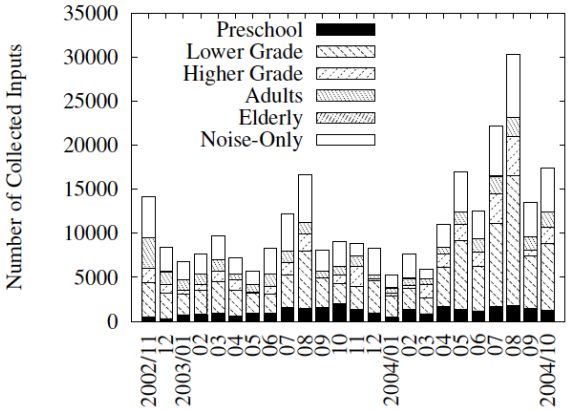

Fig. 2 Number of transcribed input utterances recorded in "Takemaru-kun" system from November 2002 to October 2004.

3. MMI Discriminative Training

3.1 MMI Objective Function

MLE training optimizes HMM model parameters that model better probability density of the observation vectors given correct word sequences. Each utterance with a known word sequence (w_r) is used to update only those HMM models that form the corresponding word sequence. On the other hand, MMI training optimizes HMM model parameters so as to maximize posterior probability of correct word sequence given the observation vectors. MMI training updates not only the corresponding HMM models, but also *all* the other models by making them more unlikely. MMI objective function is given by:

$$\mathcal{F}_{MMI}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda(O_r|M_{w_r})^k P(w_r)^k}{\sum_{\hat{w}} p_\lambda(O_r|M_{\hat{w}})^k P(\hat{w})^k} \tag{1}$$

where M_w shows the composite model of word sequence w and $P(w)$ represents the probability of this sequence from language model. Symbol λ represents HMM model parameters. Maximization of equation 1 consists of increasing the numerator term (identical to MLE function $p_\lambda(O_r|M_{w_r})$), and simultaneously decreasing the denominator term. For simplicity we represent denominator term as:

$$p_\lambda(O_r|M_{den}) = \sum_{\hat{w}} p_\lambda(O_r|M_{\hat{w}})P(\hat{w}), \tag{2}$$

with M_{den} denoting the full acoustic and language model used in recognition. However, calculating $p_\lambda(O_r|M_{den})$ over full possible word sequences has a very high computation cost. For this reason we consider only the most likely word sequences during training.

3.2 Lattice generation

One way to extract the best hypotheses mentioned previously is through generation of lattices in decoding of training utterances. From the generated lattices we can calculate $p_\lambda(O_r|M_{den})$. This method of MMI training is called lattice based MMI training and is one of the most used techniques of discriminative training. By adopting this process in MMI training we add new conditions that must be chosen only heuristically. Two of these conditions are language model and its scale factor used in generation of lattices. The language model itself is related mostly to generalization of learning process. Weak language model means less risk of overtraining the acoustic model. Language scale factor is also considered to have similar effects in MMI training. Conventional implementations of MMI generate lattices only once, and then update HMM model parameters through many iterations. This process is employed in this investigation.

3.3 Parameter Updates

An efficient way of optimizing MMI objective function is the Extended Baum-Welch algorithm. The update form of mean and variance of this algorithm, as explained in[8),11)], are:

$$\hat{\mu}_{jm} = \frac{\{\theta_{jm}^{num}(O) - \theta_{jm}^{den}(O)\} + D\mu_{jm}}{\{\gamma_{jm}^{num} - \gamma_{jm}^{den}\} + D}, \tag{3}$$

$$\hat{\sigma}_{jm}^2 = \frac{\{\theta_{jm}^{num}(O^2) - \theta_{jm}^{den}(O^2)\} + D(\mu_{jm}^2 + \sigma_{jm}^2)}{\{\gamma_{jm}^{num} - \gamma_{jm}^{den}\} + D} - \hat{\mu}_{jm}^2. \tag{4}$$

where γ_{jm} refers to Gaussian occupancy of component mixture m in state j. $\theta(O)$ and $\theta(O^2)$ represent first order and second order sufficient statistics, which are the sum of observation data and squared data respectively, weighted by posterior probability. Parameter D is inserted in these equations in order to guarantee that the updated mean and variance $(\hat{\mu}, \hat{\sigma}^2)$ will lead to local optimum of objective function 1. Furthermore this parameter is responsible for learning rate of the algorithm. If D is set too large then training is slow, however if it is set too small we may not have an increase of objective function for every iteration.[11)] shows that by setting this parameter on a per-Gaussian level to:

$$D = \max(E\gamma_{jm}^{den}, 2D_{min}). \tag{5}$$

we can have more robust learning rate. D_{min} value makes sure that we always will have a positive definite variance matrix (Σ), and parameter E has a global value which is determined heuristically from case to case.

Parameter k in 1 is the probability scale. Following[8),11)],value of k is set to

the inverse of language scale factor used in lattice generation. This parameter is usually referred as acoustic scale factor of MMI training, and is responsible for leading to a good test-set performance.

In order to keep HMM models from over fitting training data, a technique called I-smoothing is usually applied to MMI training (as in[8),11)]). This method introduces a hyper parameter τ that is selected heuristically and is responsible for adjusting the interpolation between MMI and ML objective functions in each Gaussian component according to the amount of training data available.

4. Evaluation

4.1 Experimental Conditions

An experimental evaluation of MMI training using the "Takemaru-kun" database was conducted in two phases. First we investigated the influence of initial conditions, such as a language model and language scale factor used in lattice generation, and the number of Gaussian mixture components in the acoustic model. Then we evaluated MMI performance by changing the acoustic scale factor and the I-smoothing τ parameter during parameter updates.

Evaluations were conducted separately for each speaker group. Test sets were created by extracting 10% of the number of utterances randomly for each group as shown in table I. A unique dictionary was created in order to have zero OOV (out of vocabulary) words for training utterances. The number of words included in this dictionary was 58K. Trigram language models were build from the "Takemaru-kun" database using only training utterances for each speaker group, which were used in LVCSR test.

We built from scratch acoustic models for each speaker group using their corresponding training utterances. All acoustic models consisted of 3-state left-to-right triphone HMMs of which each state output probability density was modeled by a GMM. The acoustic feature vector was a 25-dimensional vector including ΔE (energy), 12 MFCC and 12 ΔMFCC.

4.2 Effect of Initial Conditions

The effect of a language model used for generating the word lattices for MMI training in the adult speaker group is shown in fig. 3. It is observed that a bigram language model yields better word accuracy than a unigram language model. This result is not consistent with [7][8]: they have reported that the use of a unigram language model makes the MMI training more general compared to a bigram language model and it tends to yield better results. This different behavior of MMI training would be explained by the properties of speech databases. It is expected that since the performance of the acoustic model in "Takemaru-kun" data is not high enough, stronger language constraints are still effective for improving quality of word lattices.

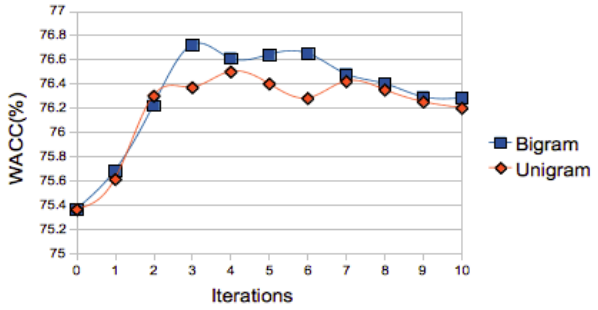

Fig. 3 Influence of language model used in lattice generation.

The effect of language scale factor in the adult speaker group is shown in fig. 4. An increase of language scale factor, i.e., a decrease of language model probabilities, gives significant improvements in word accuracy. This setting is effective for including more acoustically competitive hypotheses in word lattices and improving the discriminability of the acoustic model.

Figure 5 shows the behavior of MMI training over the different number of

Table 1 Training and Test Sets

Group	Training (Number of utterances/ time)	Test (Number of utterances)
Adult	21378/10.4 h	2375
Elderly	531/ 0.29 h	58
Lower grade	72749/ 42.2 h	8083
Preschool	16134/ 9 h	1792
Higher grade	22411/ 12 h	2490

Gaussian mixture components in the adult speaker group. A larger number of mixture components increases the risk of overtraining the acoustic model. In case of "Takemaru-kun" database, the best performance is yielded by the acoustic model with 32 Gaussian mixture components.

These results have also been observed in the other speaker groups.

4.3 Effect of Training Parameters

Results of changing I-smoothing τ and the acoustic scale factor to be optimized heuristically in the adult speaker group are shown in fig. 6 and fig. 7, respectively. In "Takemaru-kun" database, these parameters do not cause significant differences in word accuracy. These results have also been observed in the other speaker groups.

Table 2 shows comparisons of word accuracy between MMI training and MLE training in every speaker group. We can observe that MMI training yields word accuracy improvements in every speaker group. It yields 2 % absolute word accuracy improvement overall. We have found that many utterances include errors of phoneme sequences in numerator lattices. They are caused by using poor performance of the acoustic model for generating phoneme sequences from word sequences while considering multiple pronunciations. It is worthwhile to revise these errors and investigate how much further improvements are yielded by MMI training.

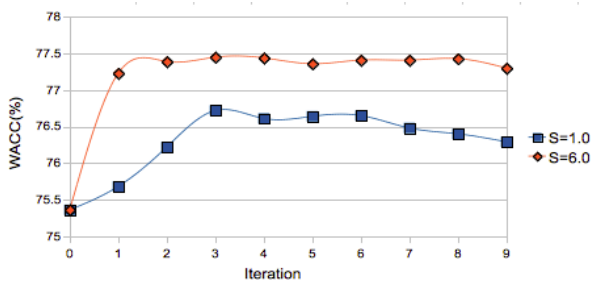

Fig. 4 Influence of language scale factor in lattice generation.

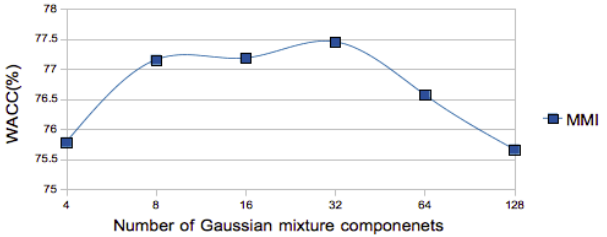

Fig. 5 Influence of number of Gaussian mixture components.

Table 2 Comparison of word accuracy between MMI and MLE.

Group	MLE acc. (%)	MMI acc. (%)
Adult	75.36	77.45
Elderly	45.16	45.62
Preschool	44.16	45.46
Lower grade	61.11	63.82
Higher grade	67.14	68.75
Total	62.30	64.54

5. Conclusions

The goal of this study was to investigate performance of MMI training using utterances recorded form a real-environment speech oriented guidance system. Training corpus was divided into five speaker groups for better observing MMI training behavior in different training conditions and different speech properties.

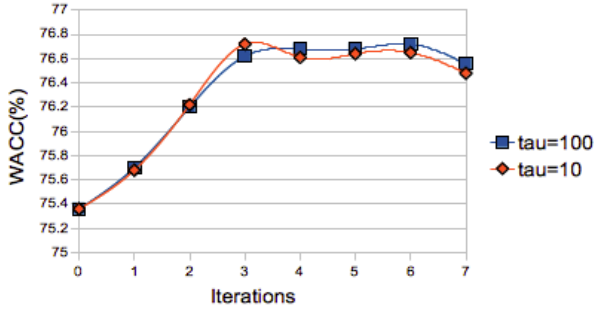

Fig. 6 Influence of I-smoothing parameter in parameter update.

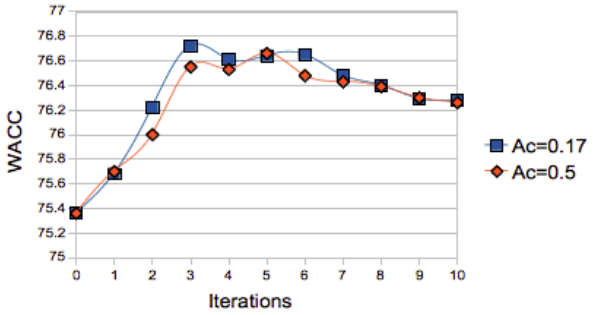

Fig. 7 Influence of acoustic scale factor in parameter update.

Results from initial training conditions showed that on speech NUI systems with a limited domain of discourse, generalization of MMI discriminative training should be kept to a minimum. Furthermore, MMI performance did not respond much to changes in acoustic scale factor and I-smoothing τ parameters, instead its performance improved considerably with changes in the quality of lattices.

Acknowledgments The authors are grateful to Dr. Erik McDermott for the invaluable advices concerning discriminative training techniques.

References

1) R. Nishimura, Y. Nishihara, R. Tsurumi, A. Lee, H. Saruwatari, and K. Shikano, "Takemaru-kun: Speech-oriented Information System for Real World Research Platform", *International Workshop on Language Understanding and Agents for Real World Interaction*, pp. 70-78, 2003.
2) J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE Trans. SAP*, vol. 2, no. 2, pp. 291-298, 1994.
3) M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition", *Computer Speech and Language*, vol. 12, no. 2, pp. 75-98, 1998.
4) M.J.F. Gales and S.J. Young, "Robust continuous speech recognition using parallel model combination", *IEEE Trans. SAP*, vol. 4, no. 5, pp. 352–359, 1996.
5) L. Deng, J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition", *IEEE Trans. SAP*, vol. 11, no. 6, pp. 568–580, 2003.
6) M.J.F. Gales, "Semi-tied covariance matrices for hidden Markov models", *IEEE Trans. on SAP*, vol. 7, no. 3, pp. 272–281, 1999.
7) S. Axelrod, V. Goel, R.A. Gopinath, P.A. Olsen and K. Visweswariah, "Subspace constrained Gaussian mixture models for speech recognition", *IEEE Trans. on SAP*, vol. 13, no. 6, pp. 1144–1160, 2005.
8) P.C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition", *Computer Speech and Language*, vol. 16, no. 1, pp. 25–47, 2002.
9) K.C. Sim and M.J.F. Gales, "Minimum phone error training of precision matrix models", *IEEE Trans. ASLP*, vol. 14, no. 3, pp. 882–889, 2006.
10) E. McDermott, T.J. Hazen, J.L. Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large-vocabulary speech recognition using minimum classification error", *IEEE Trans. on ASLP*, vol. 15, no. 1, pp. 203–223, 2007.
11) Daniel Povey "Discriminative Training for Large Vocabulary Speech Recognition" *PhD thesis, University of Cambridge*, July 2004.