Discriminative Training of Stochastic Segment Model Based on HMM Segmentation for Continuous Speech Recognition

*Yong-Joo Chung, and **Chong-Kwan Un

Abstract

In this paper, we propose a discriminative training algorithm for the stochastic segment model (SSM) in continuous speech recognition. As the SSM is usually trained by maximum likelihood estimation (MLE), a discriminative training algorithm is required to improve the recognition performance. Since the SSM does not assume the conditional independence of observation sequence as is done in hidden Markov models (HMMs), the search space for decoding an unknown input utterance is increased considerably. To reduce the computational complexity and storage amount in an iterative training algorithm for discriminative SSMs, a hybrid architecture of SSMs and HMMs is proposed. In the method, the segment (phoneme) boundaries of N-best candidate sentences are obtained by a dynamic programming using HMMs. Given the segment boundaries, the parameters of the SSM are discriminatively trained by the minimum error classification criterion based on a generalized probabilistic descent (GPD) method. With the discriminative training of the SSM, the word error rate is reduced by 17% compared with the MLE-trained SSM in speaker-independent continuous speech recognition.

I. Introduction

HMM is now the most widely used tool for speech recognition. In most HMMs, observations in a state are assumed to be generated independently and identically distributed. The assumption ignores the correlations that exist between the frames of speech signal. It is well-known that speech recognition performance is enhanced by utilizing the frame-correlation information [1].

A stochastic segment model (SSM) is one of the approaches that have been proposed to better model the spectral /temporal structure over the duration of a phoneme [2] [3]. The SSM deals with speech signal on a segment (phoneme) level rather than on a frame level.

SSMs are usually trained by maximum likelihood(ML) criterion. Maximum likelihood estimation (MLE) training maximizes the probability of observing training data given a model corresponding to the data. But, it can not usually produce an optimal model that achieves minimum classification error rate in real environment due to model incorrectness and insufficient training data and so on. In particular, enough observations of phonemes are required in the training data to reliably estimate the parameters of the SSM since it is modeled on a phoneme level. Thus, it is difficult to obtain optimal models by the MLE training of the SSM.

In this paper, a discriminative training algorithm based on a GPD method [4] is suggested for the SSM. For the discriminative training of the SSM, one needs to have an iterative algorithm that first segments input utterances by dynamic programming, and then updates the parameters of the SSM to minimize the cost function for minimum error classification. But, unlike HMMs, SSMs do not assume the conditional independence of observation sequence and this increases the search space for decoding considerably. Particularly, for continuous speech recognition, the amount of memory required as well as computation time is considerably increased in a network search algorithm for finding N-best candidate sentence hypothesis [5].

To overcome the problems, we propose a discriminative training method for the SSM by using HMM-based segmentation. In the method, phoneme boundaries of N-best candidate sentences are obtained by the HMM-based Viterbi decoding. Using the given phoneme boundaries, the parameters of the SSM are discriminatively trained by the GPD method.

This paper is organized as follows. In Section 2, we review and discuss modeling of the phonetic segments. In Section 3, the proposed scheme on discriminative training of the SSM is described. Section 4 presents our experimental

*Switching Division LG Information & Communications Ltd.
**Communications Research Laboratory Department of Electrical Engineering Korea Advanced Institute of Science and Technology

Manuscript Received 96. 8. 20

results on speaker-independent continuous speech recognition with the SSM. Finally, in Section 5, conclusions are given.

Ⅱ. Stochastic segment model

For the training of the SSM, it is first required to have phonetic segments each of which consists of a sequence of speech frames corresponding to a phonetic category. The phonetic segments may be obtained from manual segmentation based on pronunciation dictionary of words or sentences, or they can be determined in an automatic training/recognition algorithm as was developed in [2]. Since, in this paper, we obtain the phonetic segments from the HMM-based Viterbi decoding, we only describe how to model the phonetic segments.

2.1 Sampling

The phonetic segments obtained from input utterances have variable length. But, it is often required to obtain fixed-length segments for successful speech recognition [2]. In this paper, we take a linear sampling approach without interpolation which has shown satisfactory results. The linear time sampling selects a fixed number of samples which are closest in time to the sampling times spaced at equal time-intervals in the original segments. If we consider an original segment of length L, $\mathbf{X}=\{\mathbf{x}_1,\cdots,\mathbf{x}_L\}$ where $\mathbf{x}_i$ is a k-dimensional vector, the sampling process results in a fixed-length segment of length M, $\mathbf{Y}=\{\mathbf{y}_1,\cdots,\mathbf{y}_M\}$ where $\mathbf{y}_i\in\{\mathbf{x}_1,\cdots,\mathbf{x}_L\}$.

2.2 Multivariate Gaussian Mixture Model

A multivariate Gaussian mixture model may be assumed for the sampled segment Y. The observation density function for a sampled phonetic segment Y may be given as

$$P(\mathbf{Y}|\alpha)=\sum_{i=1}^{I} c_{i\alpha}\, N(\mathbf{Y},\mu_{i\alpha},\Sigma_{i\alpha}) \tag{1}$$

where α represents a phoneme and $c_{i\alpha}$ is a mixture weight. $N(\mathbf{Y},\mu_{i\alpha},\Sigma_{i\alpha})$ is a Gaussian density with a mean vector $\mu_{i\alpha}$ and a covariance matrix $\Sigma_{i\alpha}$ corresponding to the i-th mixture of the phoneme α and I represents the number of mixtures for a phonetic segment. The dimension of Y is $k\times M$ which is very large for moderate sizes of k and M. This large dimension makes it difficult to reliably estimate the full covariance matrix due to insufficient training data. Thus, it is usually assumed that samples in a segment Y are independent of each other. Although the above assumption simplifies the estimation of the covariance matrix, it does not fully take advantage of the merit of segment structure which models correlation between samples. In order to compensate for the performance degradation which may result from the simplification, we can use an additional representation which contains the information on dynamics of the sampled segment Y. This can be represented as $\hat{\mathbf{Y}}=\{\hat{\mathbf{y}}_1,\cdots,\hat{\mathbf{y}}_{M-1}\}$ where $\hat{\mathbf{y}}_i=\mathbf{y}_{i+1}-\mathbf{y}_i$. In this paper, we use diagonal covariance matrices for the samples in a phonetic segment since it may be difficult to reliably estimate the full covariance matrices of a mixture Gaussian density function with limited training data.

Ⅲ. Discriminative training of SSM

For the discriminative training of the SSM based on the GPD method, an iterative algorithm that segments each input utterance in the training set conditioned on the assumed models is required. Here one should note that an automatic segmentation algorithm by the SSM can have a problem.

For each time t, the score of each phoneme model must be computed over the possible phoneme duration. Thus, the computational complexity is increased proportionally to the product of the number of phoneme models and the maximum duration of phonemes. Particularly, for continuous speech recognition using a grammar (e.g., finite state network), the accumulated likelihood and durational information of the best path to a node corresponding to a phoneme in the finite state network must be preserved for a length in frames equal to the maximum duration of phonemes. This requires considerable increase in storage amount compared to the HMM-based search procedure where the temporal informations are needed only until the next frame. Thus, to lessen the burden of the computational complexity and storage amount, we introduce a discriminative training method using HMM-based segmentation in this section.

3.1 System overview

A block diagram of the overall recognition system with discriminative training of the SSM using HMM-based segmentation is shown in Fig. 1. The discriminative training procedure can be divided into two parts. In the first part, the HMM is used to generate segmentation for N-best candidate sentences. We can efficiently find not only the N-best lists but also the segmentation for their phonetic segments by using a frame synchronous network search algorithm [5]. In the second part, the phonetic segments are rescored by the SSM and the SSM parameters

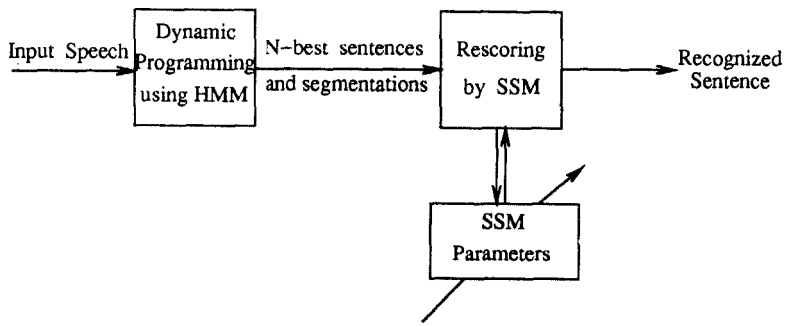

Figure 1. Block diagram of the overall recognition system with discriminative training of SSM.

are adapted by a gradient descent method.

The score of an input utterance X, given a sentence hypothesis which consists of a phoneme sequence $\alpha = \{\alpha_1, \cdots, \alpha_P\}$ of length P, is expressed using a modified likelihood function as

$$S(\mathrm{X}|\alpha) = S(\mathrm{Y}|\alpha) = \sum_{p=1}^{P} \{\ln[P(\mathrm{Y}_p|\alpha_p)] \cdot L(p) + C\} \qquad (2)$$

where Y represents a sequence of phonetic segments Y_p, and $S(\mathrm{Y}|\alpha)$ is written on the assumption the Y_p's are independent of each other. $L(p)$ is the length of p-th phonetic segment before sampling, and C is used to control phoneme insertion rate. By weighting the log-likelihood score by $L(p)$, we can take into account the length of a phonetic segment before sampling. This is necessary to prevent from favoring larger segments, which will have a higher probability per input utterance.

Since the N-best lists are available, it is possible to compute the SSM score for each hypothesis. The recognized sentence is selected as the sentence among the lists which has the highest SSM score. In our proposed architecture, the time-consuming part for finding the segmentation information is replaced by the relatively simple HMM-based Viterbi decoding process which also requires much less storage amount. Many researchers used the HMM-based segmentation information as a reliable substitute for elaborate manual segmentation results. From this fact, it is believed that possible incorrect segmentations for phonetic segments does not much worsen recognition performance.

3.2 Discriminative training of SSM

In this section, a discriminative training method based on the GPD algorithm is described for the SSM. To formulate an objective function of the discriminative training which approximates the number of error counts, two kinds of functions are defined. They are a discriminant function and a misclassification measure. For isolated word recognition, given an input utterance, a discriminant function g_k may be defined for each word k, $k = 1, \cdots, V$ where V is the number of words in the vocabulary. For continuous speech recognition, the discriminant function g_k can be defined for each candidate sentence. The candidate sentences are restricted to the N-best lists as shown in Fig. 1. The discriminant function for k-th sentence is defined using the likelihood score of (2) as

$$g_k(\mathrm{Y}^k, \Lambda_k) = S(\mathrm{Y}^k|\alpha^k) = \sum_{p=1}^{P} \{\ln[P(\mathrm{Y}_p^k|\alpha_p^k)] \cdot L(p) + C\} \qquad (3)$$

$$= \sum_{p=1}^{P} \{\ln\left[\sum_{i=1}^{I} c_{i\alpha_p^k} \cdot \prod_{m=1}^{M} N(\mu_{i\alpha_p^k m}, \Sigma_{i\alpha_p^k m})\right] \cdot L(p) + C\} \qquad (4)$$

where the index p represents a phonetic segment in the k-th candidate sentence and the superscript k on Y implies that the sampled segments of an input utterance depends on the corresponding sequence of phonemes α^k of each candidate sentence. Λ_k represents the mixture weights, means and covariances of the SSM associated with each sentence. The recognizer classifies an input utterance to c-th candidate sentence if $c = \mathrm{argmax}_l\, g_l(\mathrm{Y}^l, \Lambda_l)$.

The misclassification measure must represent the degree of misclassification of a recognizer. If an input utterance is given, a misclassification measure can be formulated as follows.

$$d_c = -g_c(\mathrm{Y}^c, \Lambda_c) + \left[\frac{1}{N-1} \sum_{i \neq c} g_i(\mathrm{Y}^i, \Lambda_i)^\eta\right]^{\frac{1}{\eta}} \qquad (5)$$

where g_c represents the SSM score for the correct sentence and η is a positive number which controls the contribution of candidate sentences in the misclassification measure. By varying the value of η, the degree of adjusting the parameters of the SSM corresponding to each candidate

sentence is determined. If we set η equal to infinity, the misclassification measure becomes

$$d_c = -g_c(Y^c, \Lambda_c) + g_b(Y^b, \Lambda_b) \tag{6}$$

where b is the index of the candidate sentence which is most probable except for the correct sentence. The meaning of misclassification measure is clearer in this form because only the most probable incorrect sentence is compared with the correct one. But possible classification error pattern may not be fully taken into account.

For minimum error classification, an objective function should be determined so that it approximately represents the number of classification error counts. Among some possibilities, an objective function which is one of smoothed zero-one cost functions of the misclassification measure is used:

$$l_c(d_c) = \frac{1}{1 + \exp(-\gamma \cdot d_c)} \tag{7}$$

where γ is a constant which is used to control the smoothness of the cost function. Since the above function emphasizes correct classification with a slight margin as well as near misses, it is robust to mismatch between training and testing data.

Discriminative training is done by minimizing the objective function for all training sentences with respect to the parameters (mixutre weights, means and covariances) of the SSM. A GPD algorithm is used to adapt the parameters. Given an input utterance, the parameters are adjusted according to

$$\Lambda_{n+1} = \Lambda_n - \varepsilon_n U \nabla l_c(d_c) \tag{8}$$

where U is positive definite matrix [6] and the step size sequence $\{\varepsilon_n\}$ satisfies i) $\sum_{n=1}^{\infty} \varepsilon_n \to \infty$ and ii) $\sum_{n=1}^{\infty} \varepsilon_n^2 < \infty$. In the adaptation, the gradient of the objective function with respect to SSM parameters is easily computed. The SSM parameters must satisfy certain constraints such as positiveness of the covariances and the requirement $\sum_i^I c_{ia} = 1$, $c_{ia} \geq 0$. We first take logarithm of the parameters and adapt it by the gradient descent. After adaptation, we take exponent of the updated parameters and finally normalize them to satisfy the constraints and meanings [6].

The proposed discriminative training algorithm of the SSM is summarized as follows.

1. The initial paprameters of the SSM are determined by a maximum likelihood estimate from the set of sampled segments which are obtained by the HMM-based Viterbi decoding on the training data.
2. For each input utterance, N-best candidate sentences and their segmentations are obtained by a dynamic programming network search algorithm based on HMM.
3. The gradients of the objective function with respect to the parameters of the SSM are computed based on the segmentaton given in step 2, and the SSM parameters are adjusted with the computed gradients.
4. If steps 2 and 3 are done for all training sentences, convergence is checked by the recognition rate on the training data. The iteration is stopped if it converges; otherwise, go to step 2.

Ⅳ. Experiments of speaker-independent continuous speech recognition

4.1 Task and baseline system

The vocabulary consists of 102 Korean words representing month, day, date and time. Many words in the vocabulary are very confusing with each other differing only in a small number of phonemes. 26 speakers uttered 20-30 sentences to construct the training and testing set. Utterances by 16 speakers were used for construction of the training data in which there are 380 sentences and 1245 words, and those by the other 10 speakers were used to form the testing data containing 231 sentences and 753 words. Each utterance was low-pass filtered with a cut-off frequency of 4.5 kHz and digitized with a sampling rate of 16 kHz. Twelve LPC cepstral coefficients plus energy for each 10 ms frame of speech were produced. In addition to these 13 coefficients, their derivatives were produced giving 26 coefficients per frame. We chose the phoneme as the basic subword unit. 28 phonemes were used in our experiment and this phoneme set is similar to the one originally used at KAIST[7].

4.2 Results of the MLE-trained SSM

In this paper, the number of samples in a phonetic segment was determined as 10, and each sample is represented by a 13-dimensional vector consisting of 12 cepstral coefficients plus energy. The samples are assumed independent and each of them is modeled by a multivariate mixture Gaussian density function with diagonal covariance matrices. In addition to the original phonetic segment, we also used another representation which considers differences between neighbouring samples.

For decoding in the experiments, we used a finite state

Table 1. Word(sentence) error rates (%) of the MLE-trained SSM ((a): Using only orginal segment representation; (b): Add a new feature representing the differences between samples.

No. of mixtures	(a)		(b)	
	Testing set	Training set	Testing set	Training set
1	29.2(67.5)	22.0(60.5)	21.2(51.5)	18.2(51.8)
2	26.8(66.6)	17.9(48.6)	19.1(48.9)	14.2(40.5)
3	26.8(62.7)	15.2(42.3)	18.3(45.8)	11.6(33.4)
4	26.1(63.2)	14.6(42.1)	18.5(47.1)	10.4(30.0)

network(FSN) grammar with a perplexity about 30. A frame synchronous network search algorithm based on HMM was used to find the N-best candidate sentences as well as their segmentations. The candidate sentences were rescored by the SSM to find the correct hypothesis based on the segmentations. We empirically determined the number of candidate sentences N, to be 5 because the correct sentence was almost always included in the top 5 candidate sentences. With N larger than 5, computation time increased without improving recognition performance.

In Table 1, we show the recognition results of the MLE-trained SSM as the number of mixture components is varied. First, the results using only the original segment representation are shown. The word (sentence) error rates (%) for the training data decrease as the number of mixture components increases. However, for the testing data, the recognition rate degrades with more than 3 mixture components. This may be due to the decrease in generalization of Gaussian mixture model for the SSM. Next, the results are shown when the representation employing the dynamics in the original segments is also used. The variation of the recognition error rates with the number of mixture components is similar to the first experiment. Comparing the two cases, we can see that remarkable performance improvement is achieved by using the dynamic representation. In particular, the word error rates are reduced by 28% and 41% for the training and testing data, respectively, when the number of mixture components is 4.

4.3 Results of discriminative training of the SSM

To improve discrimination of the SSM, a discriminative training for minimum error classification based on the GPD method was performed. The initial SSM was obtained from the MLE-training. The result is shown in Table 2. The adaptation of the parameters of the SSM

Table 2. Word(sentence) error rates (%) of the discriminatively trained SSM.

No. of mixtures	Testing data	Training data
1	16.3(41.9)	5.3(15.7)
2	16.0(40.2)	3.4(9.2)
3	15.1(35.9)	2.3(6.0)
4	15.6(37.6)	1.9(5.0)
MLE-trained SSM	18.5(47.1)	10.4(30.0)

was done after each training sentence. The incorrect sentence with the best likelihood score was only considered for the misclassification measure in (6). In the GPD method, we used ε_n as

$$\varepsilon_n = \varepsilon_0 (1 - \frac{n}{T}) \tag{9}$$

where T is the value obtained by multiplication of the maximum number of iterations and the number of whole training sentences. ε_0 is a small positive number and n is increased by 1 after each sentence. One iteration is for the whole training data set which consists of 380 sentences. The maximum number of iteration was fixed at 10.

With the discriminative training, we could observe improved recognition rates compared with the MLE-trained SSM in Table 1. Especially, the improvement is significant for the training data, reducing the word (sentence) error by 81% (83%). For the testing data, the best result was obtained when the number of mixutre components was 3 and the word (senteene) error rate was reduced by 17.0% (21.6%). The less improvement in the testing data may be due to characteristic of discriminative training which specifically increases recognition rate on training data. Also, since SSMs deal with speech signal on a phoneme level, the insufficient training data may degrade the generalization of discriminatively trained models. Although not reported in this paper, we also experimented using the misclassification measure in (5) instead of (6) to take into account multiple candidate sentences. But, the improvement was marginal, possibly due to the fact that we employ a cost function in (7) which is robust to mismatch between training and testing data.

V. Conclusions

In this paper, we proposed a discriminative training algorithm to improve the recognition performance of an MLE-trained SSM. In the method, a hybrid architecture

of SSMs and HMMs was employed to reduce the computational complexity and storage amount in an iterative training algorithm for discriminative SSMs. The likelihood scores of the SSM were obtained based on the segmentation information from the HMM. In order to model the correlation between samples in a phonetic segment, the differences between adjacent samples were used also as features. The use of the additional features improved the recognition rates considerably. For discriminative SSMs, a discriminative training algorithm based on the GPD method was performed using the N-best candidate sentences obtained from an HMM-based network search procedure. With the discriminative training of the SSM, the recognition rate was increased significantly compared with the MLE-trained SSM.

References

1. S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol.34, pp.52-59, 1986.
2. M. Ostendorf and S. Roucos, "A stochastic segment model for phoneme-based continuous speech recognition," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol.37, pp. 1857-1869, Oct.1989.
3. S. Roucos, M. Ostendorf, H. Gish, and A. Derr, "Stochastic segment modeling using the estimate-maximize algorithm," in *Proc. of IEEE Int. Conf. Acoust., Speech, Signal Processing*, New York, NY, pp.127-130, Apr.1988.
4. B. H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing*, vol.40, pp.3043-3054, Dec.1992.
5. C. H. Lee and L. R. Rabiner, "A frame-synchronous network search algorithm for connected word recognition," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol.37, pp.1649-1658, Nov.1989.
6. W. Chou, B. H. Juang, and C. H. Lee, "Segmental GPD training of HMM based speech recognizer," in *Proc. of IEEE Int. Conf. Acoust., Speech, Signal Processing*, San Francisco, CA, pp.473-476, Mar. 1992.
7. I. J. Choi, "On the Development of a Large-Vocabulary Continuous Speech Recognition System for the Korean Language," *The J. of Acoustical Society of Korea*, vol.14, no.5, pp.44-50, Oct. 1995.

▲Yong-Joo Chung

Yong Joo Chung received the B.S. degree in electronics engineering from Seoul National University in 1988 and the M.S. and Ph.D degrees from KAIST in 1990 and 1995, respectively. He has been a research engineer at LG Information and Communications, Ltd. since 1995. His current research interest includes speech recognition, neural networks and signal processing.

▲Chong-Kwan Un

Chong Kwan Un (S'63-M'64-SM'81'-F'87) was born in Seoul, Korea. He received the B.S., M.S., and Ph.D. degrees in eletrical engineering from the University of Delaware, Newark, in 1964, 1996, and 1969, respectively.

From 1969 to 1973, he was an Assistant Professor of Electrical Engineering at the University of Maine, Portland, where he taught communications and did research on synchronization problems. In May 1973 he joined the staff of the Telecommunication Sciences Center, SRI International, Menlo Park, CA, where he did research on voice digitization and bandwidth compression systems. Since June 1977 he has been with Korea Advanced Institute of Science and Technology (KAIST), where he is a Professor of Electrical Engineering, teaching and doing research in the areas of digital communications and signal processing. So far, he has supervised 59 Ph.D. and more that 110 M.S. graduates. He has authored or coauthored over 350 papers on B-ISDN, protocol design and analysis, very high-speed packet communication systems, speech coding and processing, adaptive signal processing and data communications. Also, he holds seven patents granted. From February 1982 to June 1983 he served as Dean of Engineering at KAIST.

Dr. Un is a Fellow of IEEE and a Fellow of the Korean Academy of Sciences and Engineering. Also, he is a Founding Member of the National Academy of Engineering of Korea. He received a number of awards, includ-

ing the 1976 Leonard G. Abraham Prize Paper Award from the IEEE Communications Society, the National Order of Merits from the Government of Korea, and Achievement Awards from the Korea Institute of Telematics and Electronics, the Korea Institute of Communication Sciences, and the Acoustical Society of Korea(ASK), Highest Achievement Award from KAIST He was President of the ASK from 1988 to 1989. He is a member of Tau Beta Pi and Eta Kappa Nu.