Large Margin Discriminative Semi-Markov Model for Phonetic Recognition

Sungwoong Kim, *Student Member, IEEE,* Sungrack Yun, *Student Member, IEEE,* and
Chang D. Yoo, *Member, IEEE*

Abstract—**This paper considers a large margin discriminative semi-Markov model (LMSMM) for phonetic recognition. The hidden Markov model (HMM) framework that is often used for phonetic recognition assumes only local statistical dependencies between adjacent observations, and it is used to predict a label for each observation without explicit phone segmentation. On the other hand, the semi-Markov model (SMM) framework allows simultaneous segmentation and labeling of sequential data based on a segment-based Markovian structure that assumes statistical dependencies among all the observations within a phone segment. For phonetic recognition which is inherently a joint segmentation and labeling problem, the SMM framework has the potential to perform better than the HMM framework at the expense of slight increase in computational complexity. The SMM framework considered in this paper is based on a non-probabilistic discriminant function that is linear in the joint feature map which attempts to capture long-range statistical dependencies among observations. The parameters of the discriminant function are estimated by a large margin learning framework for structured prediction. The parameter estimation problem in hand leads to an optimization problem with many margin constraints, and this constrained optimization problem is solved using a stochastic gradient descent algorithm. The proposed LMSMM outperformed the large margin discriminative HMM in the TIMIT phonetic recognition task.**

EDICS Category: SPE-RECO

I. Introduction

In automatic speech recognition (ASR), a continuous-density hidden Markov model (HMM) which is considered as a probabilistic generative model has been popularly used. A generative model represents the joint probability of the observation and label sequences, and by the Bayes rule, it is used to compute the posterior probability of the label sequence given the observation sequence. For tractable inferences (often by dynamic programming), conditional independencies among observations are incorporated into the generative model for sequential labeling task such as the ASR that has an exponentially large number of possible label sequences to consider. A generative HMM for ASR specifically imposes a frame-based Markovian structure on the label sequence in addition to the conditional independencies on the observation sequence. But, a generative HMM is limited in capturing long-range statistical dependencies, and to overcome this limitation it must use multiple overlapping features across frames. For example, the distribution of the state duration of a generative HMM is restricted to a geometric form parameterized by the self-transition probability, even though it is inconsistent with the actual duration distribution. A generative HMM is further limited in that the HMM parameters estimated by maximizing the joint probability do not lead to minimum prediction error rate. This has led to interest in discriminatively trained generative HMMs and discriminative HMMs.

Various discriminative training (DT) algorithms have been proposed to train generative HMMs. Conventional DT algorithms include the maximum mutual information (MMI) [1], minimum classification error (MCE) [2] and minimum word/phone error (MWE/MPE) [3]. The MMI maximizes an approximate posterior probability while the MCE, MWE and MPE approximately minimize the string error rate, word error rate and phone error rate on the training data, respectively. These DT algorithms, however, are liable to the over-fitting problem when the number of parameters is relatively large in comparison to the number of training data. For better generalization, recent DT algorithms have directly tried to increase the margin between the logarithm of the joint probability of the correct label sequence and that of a competing label sequence by adopting the large margin learning framework of a support vector machine (SVM) [4]–[8]. The large margin estimation (LME) [4], [6] defines a criterion to maximize the minimum positive margin among the correct label sequences. On the other hand, the soft margin estimation (SME) [5], soft large margin estimation (SLME) [9] and large-margin MCE (LM-MCE) [7], [8] consider both the incorrect label sequences and the correct label sequences by minimizing the weighted sum of the empirical risk and a generalization term which is associated with the margin.

Although the objective functions are similar, the motivations behind the SME, SLME and LM-MCE are different. The SME is motivated from the generalization bound of the classifier in statistical learning theory [10] by minimizing the error risk for the training data and simultaneously maximizing a user-defined soft margin. In [11], it has been shown that the SME improves the performances over the MCE on the mid-sized vocabulary continuous speech recognition (CSR) (5k-word Wall Street Journal) task [11]. The SLME is based on a variant of the soft margin SVM, and the performance improvement over the MCE on the small vocabulary CSR (TIDIGITS) task has been shown in [9]. In contrast to the SME and SLME, the LM-MCE is an extension of the MCE by incorporating

Manuscript received April 9, 2010; revised September 14, 2010; accepted January 3, 2011. This work was supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Culture Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2010.

The authors are with Department of EE, Korea Advanced Institute of Science and Technology, 373-1 Guseong Dong, Yuseong Gu, Daejeon 305-701, South Korea (Phone + 82-42-350-5470, Fax + 82-42-350-8590, E-mail : sungwoong.kim01@gmail.com, yunsungrack@kaist.ac.kr, cdyoo@ee.kaist.ac.kr)

the discriminative large margin in the sigmoid-loss function and is the only large-margin DT algorithm that has performed better than the MCE in the large vocabulary CSR (LVCSR) (120k-vocabulary telephony CSR) task.

Even though discriminatively trained generative HMMs have been shown to perform better than generatively trained generative HMMs in terms of prediction accuracy, these are limited to modeling local statistical dependencies using a frame-based Markovian structure in addition to assuming conditional independencies on the observation sequence. To overcome these limitations, discriminative HMMs have been applied to ASR. While generative HMMs represent the joint probability, discriminative HMMs either define a non-probabilistic discriminant function or directly represent the posterior probability. Sha *et al.* [12], [13] defined a non-probabilistic discriminant function based on the unnormalized Gaussian distributions and the HMM framework. As a side note, the authors also propose a large margin learning algorithm with a soft-max approximation. Gunawardana *et al.* [14], Sung *et al.* [15] and Morris *et al.* [16] directly model the posterior probability as an exponential distribution by HMM-like conditional random fields (CRFs). A discriminative models such as the CRF can relieve the restriction to incorporate long-range statistical dependencies in nature, since it does not assume conditional independencies on observations and allows for multiple interacting features [17]. However, all the aforementioned discriminative HMMs for ASR still impose frame-based Markovian structures in addition to conditional independencies on the observation sequence.

While most HMMs considered in the past assume only local statistical dependencies between adjacent observations and predict a label for each observation without explicit segmentation, the semi-Markov model (SMM) allows simultaneous segmentation and labeling of sequential data with a segment-based Markovian structure [18], [19]. ASR is inherently a joint segmentation and labeling problem. In comparison with the HMM framework, the SMM framework has the extended capability to use a richer class of segmental features defined over segment boundaries. Therefore, the SMM framework has the potential to perform better than the HMM framework for ASR.

Several forms of SMMs and segment models have been proposed, including the explicit duration HMM [20]–[22], the stochastic segment model [23], [24], the polynomial trajectory segment model [25], the linear trajectory model [26], [27], the nonstationary-state HMM [28] and the segmental HMM [29], [30]. However, these models have not fully exploited the benefits of a SMM. Almost all previous efforts to adopt the SMM framework have been devoted to either the incorporation of an explicit duration model into a generative HMM framework or the modeling of feature dynamics within a given segment by trajectory models under a frame-based Markovian structure[1]. In other words, in the past, the frame-based observations within a segment are assumed to follow a Markov process (frame-based Markovian assumption); frame-based observations within a segment are assumed either to be conditionally independent given both the segment length and label or to follow a Markov process. All SMMs considered in the past are generative in nature, and the improvements obtained by the previous generative SMMs over the generative HMMs were only marginal [21], [22], [32]–[35] while the performances of HMMs have been much improved by recent discriminative training methods and discriminative models [1]–[5], [7]–[9], [11]–[16].

For other tasks such as activity recognition and natural language processing [36]–[38], discriminative SMMs have been shown to perform better than discriminative HMMs. However, in the speech recognition community, a discriminative SMM has not been explored extensively.

In this paper, we propose a large margin discriminative SMM (LMSMM) for phonetic recognition. In the task of phonetic recognition, a sequence of phonetic labels must be obtained from a speech utterance without any given segmentation information. SMM is capable of simultaneously performing phonetic segmentation and labeling with segment-based features. *The contribution of this paper is that this is the first study on large margin discriminative model under the SMM framework for phonetic recognition*[2]. In contrast to what were proposed using the semi-Markov CRFs [38], [40], we define not a posterior probability but an explicit discriminant function and estimate the function parameters by structured SVM (SSVM) [41] which is a large margin learning framework for structured prediction. The proposed discriminant function is linear in the segment-based joint feature map which consists of the transition feature function, duration feature function and content feature function. The function parameters are estimated, such that the SSVM increases the score margin obtained from the discriminant function by scaling it with a loss function. This estimation process offers better generalization ability than other learning criteria for structured prediction [10], [42]. The parameter estimation problem leads to an optimization problem with many margin constraints. The stochastic gradient descent [43] with both the hard-max and the soft-max margins [12], [13] is used to solve the optimization problem of SSVM in the primal domain, since it leads to fast convergence and can handle a large number of margin constraints. Experimental results based on the TIMIT phonetic recognition show that the proposed LMSMM outperforms the large margin discriminative HMM (LMHMM) [12], [13].

The rest of the paper is organized as follows. Section II presents the proposed discriminative SMM for phonetic recognition. Section III describes the large margin training for the discriminative SMM based on the SSVM and the stochastic gradient descent algorithm. A number of experimental and comparative results are presented and discussed in Section IV, followed by a conclusion in Section V.

[1] Separate from the HMM and the SMM, hidden dynamic models have been proposed as the super-segmental models with multi-level hidden dynamic variables to capture the long-term correlation on the entire sequence based on the physical properties of speech generation [31]

[2] A preliminary version of this paper has been published at [39].

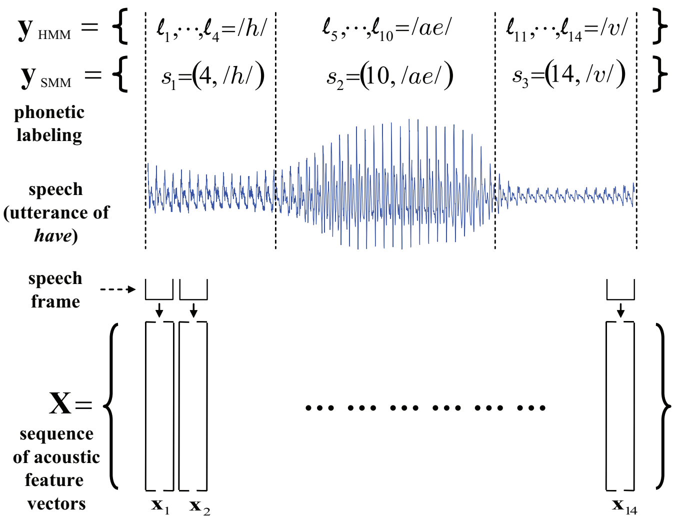
$\mathbf{y}_{\text{HMM}} = \{$
$\ell_1,\cdots,\ell_4=/h/$
$\ell_5,\cdots,\ell_{10}=/ae/$
$\ell_{11},\cdots,\ell_{14}=/v/$ $\}$
$\mathbf{y}_{\text{SMM}} = \{$
$s_1=(4,/h/)$
$s_2=(10,/ae/)$
$s_3=(14,/v/)$ $\}$
phonetic labeling
speech (utterance of *have*)
speech frame
$\mathbf{X}=$
sequence of acoustic feature vectors
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_{14}$

Fig. 1. A phonetic recognition example based on one-state monophone model. Given an utterance of "*have*", the acoustic feature vector $\mathbf{x}_t$ is extracted from the t-th speech frame, and $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_{14}\}$. Phonetic recognition of $\mathbf{X}$ yields $\mathbf{y}$ where under the HMM framework, $\mathbf{y} = \{/h/, ..., /h/, /ae/, ..., /ae/, /v/, ..., /v/\}$ while under the SMM framework, $\mathbf{y} = \{(4, /h/), (10, /ae/), (14, /v/)\}$ which means the phone *h* is in the first segment and ending at the fourth speech frame, and so others.

II. Discriminative Semi-Markov Model for Phonetic Recognition

Phonetic recognition transcribes an utterance into a sequence of phonetic labels with their position. Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{L}$ be the space of the acoustic feature vector sequences, phonetic label sequences and phonetic labels, respectively. The phonetic recognizer predicts a phonetic label sequence $\hat{\mathbf{y}}(\in \mathcal{Y})$, given a sequence of D-dimensional acoustic feature vectors $\mathbf{X}(\in \mathcal{X}) = \{\mathbf{x}_t(\in \mathbb{R}^D)\}_{t=1}^{T}$ which is extracted from a speech having a length of T frames, such that

$$\hat{\mathbf{y}} = \underset{\mathbf{y}\in\mathcal{Y}}{\operatorname{argmax}} F(\mathbf{X}, \mathbf{y}; \mathbf{w}) \tag{1}$$

where $F : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ is the discriminant function that assigns a score to every paired input and output sequence, and $\mathbf{w} \in \mathbb{R}^M$ is an M-dimensional parameter vector. An example of the phonetic recognition based on one-state monophone model is shown in Fig. 1. Given an utterance of "*have*", the acoustic feature vectors $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_{14}\}$ are extracted from all $T = 14$ frames. Then, the phonetic recognizer finds a sequence of phonetic labels $\mathbf{y}$ which maximizes $F(\mathbf{X}, \mathbf{y}; \mathbf{w})$. Here, the definition of output sequence $\mathbf{y}$ is different according to whether we use a HMM or SMM framework. In describing multi-state HMM, phonetic labels in one-state HMM correspond to state labels in multi-state HMM, and each frame is assigned to exactly one hidden state in both models.

We assume that F is a linear discriminant function as

$$F(\mathbf{X}, \mathbf{y}; \mathbf{w}) = \langle \mathbf{w}, \Phi(\mathbf{X}, \mathbf{y}) \rangle \tag{2}$$

where $\Phi(\mathbf{X}, \mathbf{y}) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^M$ is the joint feature map which maps a paired input and output sequence into an M-dimensional feature space to characterize the statistical dependencies on *input and output pairs*. Discriminant function can either be defined non-probabilistically or be derived probabilistically by directly modeling the posterior probability $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$. When modeling the posterior distribution by a member of the exponential family and decoding based on the maximum a posteriori criterion, $F(\mathbf{X}, \mathbf{y}; \mathbf{w})$ and $\Phi(\mathbf{X}, \mathbf{y})$ should be a function of $\log p(\mathbf{y}|\mathbf{X}, \mathbf{w})$ and a function of sufficient statistics, respectively.

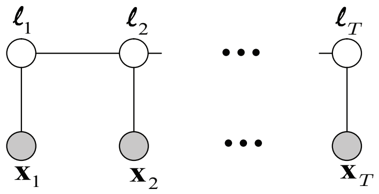

Fig. 2. An undirected graph of discriminative HMM.

The inference problem for phonetic recognition is to find the optimal label sequence, $\operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} \langle \mathbf{w}, \Phi(\mathbf{X}, \mathbf{y}) \rangle$, given $\mathbf{X}$ and $\mathbf{w}$. Note that if we define $\mathbf{y} = \{\ell_1, \ell_2 ..., \ell_T\}$ where $\ell_t(\in \mathcal{L})$ is the phonetic label of the t-th frame, the number of possible $\mathbf{y}$ grows as $O(|\mathcal{L}|^T)$. This combinatorial explosion makes inferences intractable. Therefore, a Markovian assumption between labels has been adopted to decompose Φ into a sum of local feature functions for tractable inferences. In the following section, we describe two discriminative Markov models for phonetic recognition: previously proposed discriminative HMM and the proposed discriminative SMM.

A. Discriminative HMM

A discriminative HMM for phonetic recognition assumes a frame-based Markovian structure and predicts a phonetic label for each observation without explicit phone segmentation. An undirected graph of discriminative HMM is illustrated in Fig. 2. Here, we assume a one-state HMM. Each observation is assigned to exactly one hidden state and one phonetic label, i.e. $\mathbf{y} = \{\ell_1, \ell_2 ..., \ell_T\}$, where $\ell_t \in \mathcal{L}$ is the phonetic label of the t-th observation. Henceforth, the terms frame and observation will be used interchangeably. For example, the correct label sequence $\mathbf{y}$ associated with the utterance of "*have*" in Fig. 1 is $\{/h/, ..., /h/, /ae/, ..., /ae/, /v/, ..., /v/\}$. Even though a graph in Fig. 2 is based on the assumption of one-state HMM, the structure of a multi-state HMM does not differ from the basic graph structure in Fig. 2 in that a phonetic label in one-state HMM corresponds to a state label in multi-state HMM and each frame is assigned to one hidden state in both models.

In the discriminative HMM, ℓ_t depends only on ℓ_{t-1}, ℓ_{t+1} and the acoustic feature vector of the t-th observation $\mathbf{x}_t$. This frame-based Markovian property decomposes a joint feature map $\Phi(\mathbf{X}, \mathbf{y})$ into a sum over frame specific features ϕ as

$$\Phi(\mathbf{X}, \mathbf{y}) = \sum_{t=1}^{T} \phi(\ell_{t-1}, \ell_t, \mathbf{x}_t) \tag{3}$$

where ϕ consists of two feature functions defined by pairs of adjacent labels and by pairs of label and acoustic feature vector.

Even though discriminative HMMs including HMM-like CRFs can originally relax the independence assumptions between adjacent observations, previous discriminative HMMs for phonetic recognition defined only frame-based local features under a graph structure shown in Fig. 2 [12]–[16]. Sha *et al.* [12], [13], Gunawardana *et al.* [14] and Sung *et al.* [15] defined local features derived from the Gaussian-mixture HMM while Morris *et al.* [16] defined local features using frame-level posterior estimates of phone and phonological attribute classes by multilayer perceptrons.

Using a frame-based Markovian property, an efficient inference algorithm, called Viterbi algorithm, for phonetic recognition is derived as follows. Let $V(t,\ell)$ be the maximal score for all partial labelings starting from 1 to t, such that the last label is ℓ. Dynamic programming can be used to carry out the following recursion

$$V(t,\ell) = \max_{\ell' \in \mathcal{L}} \Big(V(t-1,\ell') + \langle \mathbf{w}, \phi(\ell',\ell,\mathbf{x}_t)\rangle \Big). \qquad (4)$$

The optimal $\mathbf{y}$ is obtained by backtracking the path corresponding to $\max_\ell V(T,\ell)$. The recursion requires the computation of $\langle \mathbf{w}, \phi\rangle$ at $O(|\mathcal{L}|^2 T)$ times.

B. Discriminative SMM

Phonetic recognition is inherently a joint segmentation and labeling problem of speech observations. In comparison with the HMM framework, the SMM framework [18], [25], [32], [40], [44], [45] provides the ability not only to label but to simultaneously segment an input sequence with segment-based rich features and therefore, has the potential to perform better for this task. In the past, the benefits of a SMM had not been fully exploited. Previously considered SMMs exploit only local statistical dependencies among observations (frame-based features) using a frame-based Markovian structure. Almost all previous efforts using SMM for ASR were limited to either the incorporation of an explicit duration model into a generative HMM framework [20]–[22] or the modeling of feature dynamics within a given segment by trajectory models under a frame-based Markovian structure [26]–[30], [46]. Thus, several studies have shown that there is virtually no performance difference between the generative SMM and the generative HMM [21], [22], [32]–[35]. On the other hand, many studies report significant performance improvement using the discriminative HMM over the generative HMM [1]–[5], [9], [11]–[13], [16]. Moreover, for other tasks such as activity recognition and natural language processing [36]–[38], the discriminative SMMs have been shown to perform better than discriminative HMMs. However, the potential of the discriminative SMM has not been explored in the speech recognition community. This motivates the study of LMSMM for phonetic recognition.

The proposed discriminative SMM for phonetic recognition defines a linear discriminant function F as in (2). An undirected graph of discriminative SMM based on one-state monophone model is shown in Fig. 3. A discriminative SMM assumes a segment-based Markovian structure and can be used for segmentation and phone label prediction. It assigns variable number of frames to a hidden state that represents a segment. Additionally, the observation behavior within a segment is non-Markovian. Thus, $\mathbf{y}$ is defined as a sequence of phonetic segments, i.e. $\mathbf{y} = \{s_1, s_2 ..., s_J\}$, where the j-th segment $s_j = (n_j, \ell_j)$. Here, n_j, ℓ_j and J denote the ending frame of the j-th segment, the phonetic label of the j-th segment and the total number of segments, respectively. For instance, the correct segment sequence $\mathbf{y}$ associated with the utterance of "*have*" in Fig. 1 is $\{(4, /h/), (10, /ae/), (14, /v/)\}$. The diagram in Fig. 4 describes a typical example of segmentation and labeling. The segment $\{\mathbf{x}_{n_{j-1}+1}, ..., \mathbf{x}_{n_j}\}$ bears the phonetic label $\ell_j \in \mathcal{L}$, and $n_0 = 0$ and $n_J = T$ (there are a total of T frames, and T is the last frame index of the J-th segment) while for all j, $n_{j+1} > n_j$. Note that the number of segments J itself is a variable.

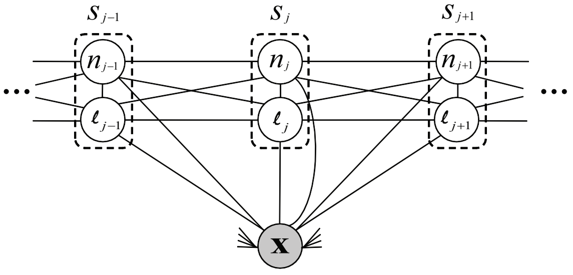

Fig. 3. An undirected graph of discriminative SMM.

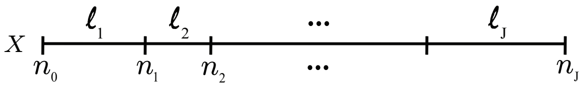

Fig. 4. A typical example of segmentation and labeling.

In discriminative SMM, s_j depends only on s_{j-1}, s_{j+1} and $\mathbf{X}$. This segment-based Markovian property decomposes a joint feature map $\Phi(\mathbf{X},\mathbf{y})$ into a sum over segment features ϕ as

$$\Phi(\mathbf{X},\mathbf{y}) = \sum_{j=1}^{J} \phi(\ell_{j-1}, \ell_j, n_{j-1}, n_j, \mathbf{X}). \qquad (5)$$

In the following subsections, detailed segment feature function and efficient inference algorithm for discriminative SMM are discussed.

C. Segment Feature Function

To capture the statistical characteristics within individual segments and between adjacent phonetic segments of variable length, we construct the segment feature function ϕ by concatenating the transition feature function ϕ^t, duration feature function ϕ^d and content feature function ϕ^c as follows

$$\begin{aligned}
&\phi(\ell_{j-1}, \ell_j, n_{j-1}, n_j, \mathbf{X}) \\
&\quad = [(\phi^t(\ell_{j-1}, \ell_j))^T, (\phi^d(\ell_j, n_{j-1}, n_j))^T, \\
&\qquad (\phi^c(\ell_j, n_{j-1}, n_j, \{\mathbf{x}_t\}_{t=n_{j-1}+1}^{n_j}))^T\ ]^T
\end{aligned} \qquad (6)$$

where components of each feature function are described below.

1) Transition Feature: Under the SMM framework, the transition feature $\phi^t_{(\ell',\ell)}$ is defined as an indicator function for phonetic transition from ℓ' to ℓ. This is shown below

$$\phi^t_{(\ell',\ell)}(\ell_{j-1}, \ell_j) = \delta(\ell_{j-1} = \ell',\ \ell_j = \ell) \tag{7}$$

where $\delta(\ell_{j-1} = \ell',\ \ell_j = \ell)$ is the Kronecker delta function that is equal to one when $\ell_{j-1} = \ell'$ and $\ell_j = \ell$ and zero otherwise. Here, the elements of ϕ^t are transition features for all pairs of phonetic labels: $\phi^t = [\phi^t_{(/h/,/h/)}, \phi^t_{(/h/,/ae/)}, ...]^T$ where $\mathcal{L} = \{/h/, /ae/, ...\}$. Transition features aim to capture statistical dependencies between two neighboring phones and are related to the bigram language model in that the weights of transition features in the discriminant function of the SMM framework ($\mathbf{w}$ in (2)) can be considered the logarithms of unnormalized transition probabilities.

2) Duration Feature: The gamma distribution is known to be a good model for the distribution of the phone durations [47], and we define the duration feature for phone ℓ, ϕ^d_ℓ, as the sufficient statistics of the gamma distribution. This is given as

$$\phi^d_\ell(\ell_j, n_{j-1}, n_j) = \begin{bmatrix} \log(n_j - n_{j-1}) \\ n_j - n_{j-1} \\ 1 \end{bmatrix} \delta(\ell_j = \ell). \tag{8}$$

The elements of ϕ^d are duration features for all phonetic labels such that $\phi^d = [(\phi^d_{/h/})^T, (\phi^d_{/ae/})^T, ...]^T$. A direct consequence of the frame-based Markovian assumption in the HMM is that phone durations have a geometric distribution defined by the probability of the self-transition. This is not adequate to model the actual phone duration distribution. On the other hand, a segment-based Markovian structure of the SMM permits an explicit duration model using the gamma distribution, which provides a suitable distribution shape for modeling the phone durations.

3) Content Feature: Content features are defined by both the labeled segment and all observations within a phone segment. In most cases, state observation probabilities of generative HMMs are Gaussian. Thus, Gaussian sufficient statistics calculated for each observation are widely used as content features of discriminative HMMs for ASR [12]–[15], [48], [49]. However, these frame-based content features are limited in capturing long-range statistical dependencies on the observations.

The discriminative SMM allows a non-Markovian behavior within a segment, and we use the averages of acoustic feature vectors within a phone segment to construct a segment-based content feature that captures long-range statistical dependencies on inputs. First, we divide a segment into a number of bins and then take averages of the Gaussian sufficient statistics of the acoustic feature vectors within each bin. Let A be a $(D+1)$-by-$(D+1)$ symmetric matrix and $\text{vec}(A)$ be the $((D+1)(D+2)/2)$-dimensional vector whose elements are from the upper triangular part of A. The content feature for the pair of the phone ℓ and the k-th bin, $\phi^c_{(\ell,k)}$, is given by[3]

$$\begin{aligned}&\phi^c_{(\ell,k)}(\ell_j, n_{j-1}, n_j, \{\mathbf{x}_t\}_{t=n_{j-1}+1}^{n_j}) \\ &\quad = \frac{B(\ell)}{n_j - n_{j-1}} \sum_{t\in b_k} \text{vec}\left(\begin{bmatrix} \mathbf{x}_t\mathbf{x}_t^T & \mathbf{x}_t \\ \mathbf{x}_t^T & 1 \end{bmatrix}\right)\delta(\ell_j = \ell)\end{aligned} \tag{9}$$

where

$$\begin{aligned} b_k \;=\; & \{n_{j-1} + \frac{n_j - n_{j-1}}{B(\ell)}(k-1) + 1, ..., \\ & n_{j-1} + \frac{n_j - n_{j-1}}{B(\ell)}k\}, \quad k \in \{1, ..., B(\ell)\}, \end{aligned} \tag{10}$$

and $B(\ell)$ denotes the number of bins according to the phonetic label ℓ. The elements of ϕ^c are content features for all pairs of phonetic label and bin: $\phi^c = [(\phi^c_{(/h/,1)})^T, (\phi^c_{(/h/,2)})^T, ..., (\phi^c_{(/ae/,1)})^T, (\phi^c_{(/ae/,2)})^T, ...]^T$. The statistical characteristics of acoustic feature vectors may vary within a segment. Thus, we divide a segment into a number of bins and assign different $\mathbf{w}$ to each bin. This is similar to modeling smooth trajectories of acoustic feature vectors by deterministic mappings[4], and bins can be regarded as sub-states [18]. In addition, the content feature in each bin, which is obtained by the averaging, becomes less sensitive to variation in acoustic feature vectors across frames. In our case, the number of frames in each bin is on average 2.6 (26ms), and the statistical characteristics of the acoustic feature vectors within a bin does not vary significantly. This idea of feature averaging is in accordance with the segmental features proposed in [45], [50]. However, there are other long-range features such as the temporal pattern (TRAP) features [51] and modulation spectrum (MS) features [52], [53]. In these approaches, temporal trajectories of spectral energies in individual critical bands over windows of upto one-second length are used as features for pattern classification where the artificial neural network is often used. In comparison to the TRAP and MS features, the advantage of the proposed content features is that under the SMM framework, it leads to a linear discriminant function which is of low computational complexity, and the linear discriminant function allows a large margin training based on the SSVM to be used.

Since the average of the Gaussian sufficient statistics in each bin is calculated and the content features for all phonetic label and bin pairs are concatenated with the Kronecker delta function, the dimension of the proposed content feature of each segment is fixed to $\frac{(D+1)(D+2)}{2}\sum_\ell B(\ell)$.

D. Initial Estimation of Parameters

The definitions of ϕ^t, ϕ^d and ϕ^c can be related to the probabilistic model in the SMM framework in that if we select $\mathbf{w}$ properly, then F in (2) is (approximately) equal to

[3] (9) is based on a single Gaussian assigned to each bin. The extended content feature pertaining to the multiple Gaussian mixtures is described in the next subsection.

[4] A segment is divided into bins which have the same lengths without a forced alignment.

$\log p(\mathbf{X}, \mathbf{y}|\mathbf{w})$. To see this, we first decompose $\log p(\mathbf{X}, \mathbf{y}|\mathbf{w})$ as follows:

$$\log p(\mathbf{X}, \mathbf{y}|\mathbf{w}) = \log p(\mathbf{y}|\mathbf{w}) + \log p(\mathbf{X}|\mathbf{y}, \mathbf{w}). \tag{11}$$

In the SMM framework, we can further decompose the first term of the right-hand side of above equation into two parts:

$$\begin{aligned} \log p(\mathbf{y}|\mathbf{w}) &= \log p(\{(n_j, \ell_j)\}_{j=1}^J|\mathbf{w}) \\ &= \log p(\{\ell_j\}_{j=1}^J|\mathbf{w}) + \log p(\{n_j\}_{j=1}^J|\{\ell_j\}_{j=1}^J, \mathbf{w}) \quad (12) \\ &= \sum_{j=1}^J \log p(\ell_j|\ell_{j-1}, \mathbf{w}) + \sum_{j=1}^J \log p(n_j - n_{j-1}|\ell_j, \mathbf{w}). \quad (13) \end{aligned}$$

Therefore, if we set the parameter associated to the transition feature $\phi^t_{(\ell',\ell)}$ as the logarithm of the transition probability from ℓ' to ℓ, i.e.

$$w^t_{(\ell',\ell)} = \log p(\ell|\ell'), \tag{14}$$

then the first term of the right-hand side of (13) becomes

$$\sum_{j=1}^J \log p(\ell_j|\ell_{j-1}, \mathbf{w}) = \langle \mathbf{w}^t, \sum_{j=1}^J \phi^t(\ell_{j-1}, \ell_j)\rangle \tag{15}$$

where $\mathbf{w}^t = [w^t_{(/h/,/h/)}, w^t_{(/h/,/ae/)}, ...]^T$. Likewise, note that we model the phone duration by the gamma distribution, i.e.

$$\begin{aligned} &\log p(n_j - n_{j-1}|\ell_j, \mathbf{w}) \\ &= (\gamma_{\ell_j} - 1)\log(n_j - n_{j-1}) - \frac{n_j - n_{j-1}}{\theta_{\ell_j}} \\ &\quad -(\gamma_{\ell_j}\log\theta_{\ell_j} + \log\Gamma(\gamma_{\ell_j})) \end{aligned} \tag{16}$$

where γ_{ℓ_j} and θ_{ℓ_j} are the shape parameter and scale parameter for phone ℓ_j, respectively. If we set

$$\mathbf{w}^d_\ell = \begin{bmatrix} \gamma_\ell - 1 \\ -\frac{1}{\theta_\ell} \\ -(\gamma_\ell \log\theta_\ell + \log\Gamma(\gamma_\ell)) \end{bmatrix}, \tag{17}$$

the second term of the right-hand side of (13) can be expressed as

$$\sum_{j=1}^J \log p(n_j - n_{j-1}|\ell_j, \mathbf{w}) = \langle \mathbf{w}^d, \sum_{j=1}^J \phi^d(\ell_j, n_{j-1}, n_j)\rangle \tag{18}$$

where $\mathbf{w}^d = [(\mathbf{w}^d_{/h/})^T, (\mathbf{w}^d_{/ae/})^T, ...]^T$.

Similarly, the conditional independencies among random variables in the SMM lead to the decomposition of the second term of the right-hand side of (11) as

$$\begin{aligned} \log p(\mathbf{X}|\mathbf{y}, \mathbf{w}) &= \sum_{j=1}^J \log p(\{\mathbf{x}_t\}_{t=n_{j-1}+1}^{n_j}|\ell_j, n_{j-1}, n_j, \mathbf{w}) \\ &= \sum_{j=1}^J \sum_{k=1}^{B(\ell_j)} \frac{B(\ell_j)}{n_j - n_{j-1}} \sum_{t\in b_k} \log p(\mathbf{x}_t|\ell_j, k, \mathbf{w}) \quad (19) \end{aligned}$$

Here, we further decompose the segment-level value into the sum of bin-level averages and use the Gaussian mixture to model the acoustic feature vectors in each bin as follows:

$$\log p(\mathbf{x}_t|\ell, k, \mathbf{w}) = \log \sum_{q=1}^Q c_{(\ell,k,q)} \mathcal{N}(\mathbf{x}_t|\mu_{(\ell,k,q)}, \Sigma_{(\ell,k,q)}) \tag{20}$$

where q, Q and c denote the mixture component, the number of mixtures and the mixture weight, respectively. To obtain a linear discriminant function, we approximate the above mixture by the single most dominant Gaussian as

$$\begin{aligned} &\log p(\mathbf{x}_t|\ell, k, \mathbf{w}) \approx \log c_{(\ell,k,q^*)} \mathcal{N}(\mathbf{x}_t|\mu_{(\ell,k,q^*)}, \Sigma_{(\ell,k,q^*)}) \quad (21) \\ &= \frac{1}{2}\left(\begin{bmatrix} -\Sigma^{-1}_{(\ell,k,q^*)} & \Sigma^{-1}_{(\ell,k,q^*)}\mu_{(\ell,k,q^*)} \\ (\mu_{(\ell,k,q^*)})^T\Sigma^{-1}_{(\ell,k,q^*)} & -c'_{(\ell,k,q^*)} \end{bmatrix} \bullet \begin{bmatrix} \mathbf{x}_t\mathbf{x}_t^T & \mathbf{x}_t \\ \mathbf{x}_t^T & 1 \end{bmatrix}\right) \quad (22) \end{aligned}$$

where $q^* = \text{argmax}_q\, c_{(\ell,k,q)} \mathcal{N}(\mathbf{x}_t|\mu_{(\ell,k,q)}, \Sigma_{(\ell,k,q)})$,

$$\begin{aligned} c'_{(\ell,k,q)} &= (\mu_{(\ell,k,q)})^T \Sigma^{-1}_{(\ell,k,q)} \mu_{(\ell,k,q)} \\ &\quad -2\log(c_{(\ell,k,q)}(2\pi)^{-\frac{D}{2}}|\Sigma_{(\ell,k,q)}|^{-\frac{1}{2}}), \end{aligned} \tag{23}$$

and $A \bullet B$ denotes the matrix inner product such that $A \bullet B = \text{tr}(A^T B)$. Note that the approximation of multiple Gaussian mixtures by the single most dominant Gaussian is performed not only once for initialization but every time the segment feature function ϕ is computed for inference and training. Here, the matrix inner product is between two symmetrical matrices; therefore, if we set the parameters of the content feature by using a reparameterization matrix of the mixture parameters as follows:

$$\mathbf{w}^c_{(\ell,k,q)} = \frac{1}{2}\widetilde{\text{vec}}\left(\begin{bmatrix} -\Sigma^{-1}_{(\ell,k,q)} & \Sigma^{-1}_{(\ell,k,q)}\mu_{(\ell,k,q)} \\ (\mu_{(\ell,k,q)})^T\Sigma^{-1}_{(\ell,k,q)} & -c'_{(\ell,k,q)} \end{bmatrix}\right) \tag{24}$$

where $\mathbf{w}^c_{(\ell,k)} = [(\mathbf{w}^c_{(\ell,k,1)})^T, ..., (\mathbf{w}^c_{(\ell,k,Q)})^T]^T$ and $\widetilde{\text{vec}}(A)$ is equal to $\text{vec}(A)$ with the off-diagonal terms multiplied by two, then,

$$\log p(\mathbf{X}|\mathbf{y}, \mathbf{w}) = \langle \mathbf{w}^c, \sum_{j=1}^J \phi^c(\ell_j, n_{j-1}, n_j, \{\mathbf{x}_t\}_{t=n_{j-1}+1}^{n_j})\rangle \tag{25}$$

where $\mathbf{w}^c = [(\mathbf{w}^c_{(/h/,1)})^T, (\mathbf{w}^c_{(/h/,2)})^T, ...]^T$. Note that in the case of multiple mixtures, we modify $\phi^c_{(\ell,k)}$ in (9) such that $\phi^c_{(\ell,k)} = [\mathbf{0}, ..., \phi^c_{(\ell,k,q*)}, ..., \mathbf{0}]^T$. Thus, if $\mathbf{w}$ is assigned according to (14), (17) and (24), the linear discriminant function F in (2) is (approximately) equal to $\log p(\mathbf{X}, \mathbf{y}|\mathbf{w})$:

$$\log p(\mathbf{X}, \mathbf{y}|\mathbf{w}) \approx \langle \mathbf{w}, \Phi(\mathbf{X}, \mathbf{y})\rangle \tag{26}$$

where $\mathbf{w} = [(\mathbf{w}^t)^T, (\mathbf{w}^d)^T, (\mathbf{w}^c)^T]^T$, and the dimension M of the feature space mapped by Φ becomes

$$M = |\mathcal{L}|^2 + |\mathcal{L}| + Q\frac{(D+1)(D+2)}{2}\sum_\ell B(\ell). \tag{27}$$

Note that in our task, the segmentation information is provided only during training while in the testing, the phonetic recognition is performed via simultaneous phonetic segmentation and labeling. TIMIT [54] provides phone segmentation information, and we used it during training (see Section IV); however, other speech corpora generally do not provide such information, and this information must be obtained either by manual segmentation or by using the Viterbi algorithm. For good starting point, we estimate initial parameters $\mathbf{w}_0$ by the maximum likelihood (ML) criterion: $\log p(\ell|\ell')$, γ, θ, c, μ and Σ are first estimated by the ML criterion with segmentation information, and then $\mathbf{w}_0$ is set by (14), (17) and (24). From $\mathbf{w}_0$, a large margin training is performed where $\mathbf{w}$ is not constrained for valid probabilities any more. However, the constraint of $\mathbf{w}^c$ to maintain positive definiteness of the matrix in (24) can be imposed for a stable performance while $\mathbf{w}$ is updated by large margin training. And this constraint is easily satisfied by the projection using eigenvector decomposition after each update [13]. Also, the most dominant component q^* is determined such that $q^* = \mathrm{argmax}_q \langle \mathbf{w}^c_{(\ell,k,q)}, \phi^c_{(\ell,k,q)} \rangle$, where $\phi^c_{(\ell,k,q)}$ is equivalent to $\phi^c_{(\ell,k)}$ in (9).

E. SMM Inference

Let $V(t,\ell)$ be the maximal score for all partial segmentations such that the last segment ends at the t-th frame with label ℓ, and let $U(t,\ell)$ be a tuple of length d and previous label ℓ' occupied by the best path where phone ℓ' transits to phone ℓ at time $t-d$. Similar to the Viterbi algorithm for the HMM inference, we can derive the recursion of the Viterbi-like dynamic programming for efficient SMM inference as

$$\begin{aligned} U(t,\ell) = & \operatorname*{argmax}_{(d,\ell')\in\{1,...,R(\ell)\}\times\mathcal{L}} \Big(V(t-d,\ell') + \\ & \langle \mathbf{w}, \phi(\ell',\ell,t-d,t,\mathbf{X}) \rangle \Big) \qquad (28) \\ V(t,\ell) = & \max_{(d,\ell')\in\{1,...,R(\ell)\}\times\mathcal{L}} \Big(V(t-d,\ell') + \\ & \langle \mathbf{w}, \phi(\ell',\ell,t-d,t,\mathbf{X}) \rangle \Big) \qquad (29) \end{aligned}$$

where $R(\ell)$ is the range of admissible durations of phone ℓ to ensure tractable inference. Once the recursion reaches the end of the sequence, we traverse $U(t,\ell)$ backwards to obtain segmentation information of the sequence. An implementation of the recursion in (28) and (29) requires $O(T|\mathcal{L}|\sum_\ell R(\ell))$ computations of $\langle \mathbf{w}, \phi \rangle$. In the task of phonetic recognition based on one-state monophone model (see Section IV), we set $\sum_\ell R(\ell) = 1280$ and $|\mathcal{L}| = 48$. Thus, if we assume that the computational complexities for calculating $\langle \mathbf{w}, \phi \rangle$ are about the same for HMM and SMM frameworks, the SMM inference requires about 26 times more computation than the HMM inference. To save computation, the maximum values in (28) and (29) are obtained by searching through not the whole search space $\{1,...,R(\ell)\}\times\mathcal{L}$ but a subspace of lower resolution - $\{1, d_\ell, 2d_\ell, ..., R(\ell)\}\times\mathcal{L}$ where $d_\ell > 1$ is the search resolution for the phone ℓ (longer-length phones have larger d_ℓ than shorter-length phones). In our implementation, the SMM inference takes about 4 times more computation than the HMM inference.

III. Large Margin Training

This section describes a method to train the discriminative SMM parameters. Given a set of training pairs $\{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^N$ where $\mathbf{y}_i$ is the sequence of phonetic segments for the i-th input $\mathbf{X}_i$, and N is the number of training pairs, the goal of training is to find $\mathbf{w}$ so that the decision criterion in (1) leads to the minimum prediction error rate on unseen data. In this paper, we use a large margin learning framework for structured prediction, SSVM [41], due to its better generalization ability than other learning criteria such as the conditional maximum likelihood by maximizing the separation margin scaled with a loss function [10], [42]. We adopt the stochastic gradient descent [43] to solve the optimization problem of SSVM due to the theoretical and experimental proofs of fast convergence and robustness in handling a large number of margin constraints. In the following, we first review SSVM, and then explain the stochastic gradient descent algorithm to solve our optimization problem.

A. Structured Support Vector Machine

The SSVM finds $\mathbf{w}$ such that the separation margin is maximized (equivalent to the minimization of the square of the magnitude of $\mathbf{w}$), and the sum of the slack variables ξ_i is minimized under the constraints that the difference between the discriminant function given $(\mathbf{X}_i, \mathbf{y}_i)$ and the discriminant function given $(\mathbf{X}_i, \mathbf{y}), \mathbf{y} \neq \mathbf{y}_i$, is at least larger than the scaled margin subtracted by the slack variable for all $i = 1, ..., N$ as follows [12], [13], [41], [55]:

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{i=1}^N \xi_i \qquad (30) \\ \text{s.t.} \quad & \langle \mathbf{w}, \Delta\Phi(\mathbf{X}_i, \mathbf{y}) \rangle \geq \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i \\ & \xi_i \geq 0, \quad \forall i, \ \forall \mathbf{y} \in \mathcal{Y} \backslash \mathbf{y}_i, \end{aligned}$$

where

$$\begin{aligned} \langle \mathbf{w}, \Delta\Phi(\mathbf{x}_i, \mathbf{y}) \rangle &= F(\mathbf{X}_i, \mathbf{y}_i; \mathbf{w}) - F(\mathbf{X}_i, \mathbf{y}; \mathbf{w}) \\ &= \langle \mathbf{w}, \Phi(\mathbf{x}_i, \mathbf{y}_i) - \Phi(\mathbf{x}_i, \mathbf{y}) \rangle, \qquad (31) \end{aligned}$$

and $C > 0$ is a constant that controls the trade-off between margin maximization and training error minimization, and $\Delta(\mathbf{y}_i, \mathbf{y})$ is a loss function which quantifies the difference between $\mathbf{y}$ and $\mathbf{y}_i$. The separation margin is scaled with a loss function so that the margin constraint with high loss is penalized much more than that with low loss. This is illustrated in Fig. 5. The discriminant functions given the correct segment sequence and other two incorrect segment sequences are denoted by circle, rectangle and triangle, respectively. Let the loss between circle and rectangle be larger than that between circle and triangle. By scaling the separation margin with a loss, the rectangle is further away from the circle than the triangle is from the circle. Thus, we reduce the risk of predicting the rectangle which has high loss.

A loss function is usually a nonnegative function with the following property: $\forall i$,

$$\begin{cases} \Delta(\mathbf{y}_i, \mathbf{y}) > 0, & \text{if } \mathbf{y} \neq \mathbf{y}_i, \\ \Delta(\mathbf{y}_i, \mathbf{y}) = 0, & \text{if } \mathbf{y} = \mathbf{y}_i. \end{cases} \qquad (32)$$

In [4], the zero-one loss function is used; however, it does not allow different penalties to be given to constraints with different loss: $\forall \mathbf{y} \in \mathcal{Y} \setminus \mathbf{y}_i$, $\Delta(\mathbf{y}_i, \mathbf{y}) = 1$. In [5], [12], [13], [55], a loss function based on the Hamming distance between $\mathbf{y}$ and $\mathbf{y}_i$ is used where the Hamming distance is defined as the number of mismatches between $\mathbf{y}$ and $\mathbf{y}_i$ at frame level. In this paper, we use a loss function based on the Hamming distance to provide greater penalty to the constraint with higher loss than that with lower loss, and the loss is defined as

$$\Delta(\mathbf{y}_i, \mathbf{y}) = \sum_{t=1}^T \delta(\ell_t^i \neq \ell_t) \qquad (33)$$

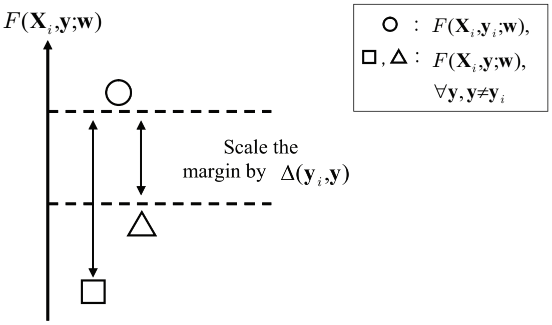
$F(\mathbf{X}_i,\mathbf{y};\mathbf{w})$
○ : $F(\mathbf{X}_i,\mathbf{y}_i;\mathbf{w})$,
□, △ : $F(\mathbf{X}_i,\mathbf{y};\mathbf{w})$, $\forall\mathbf{y},\mathbf{y}\neq\mathbf{y}_i$
Scale the margin by $\Delta(\mathbf{y}_i,\mathbf{y})$

Fig. 5. The circle, rectangle and triangle denote the discriminant function given the correct segment sequence and the other two incorrect segment sequences, respectively. By scaling the margin, the rectangle which has a high loss is further away from the circle than the triangle which has a low loss is from the circle.

where ℓ_t^i is the phonetic label of the t-th frame of $\mathbf{y}_i$. Even though the string-based phone error rate by edit distances is a more appropriate measure for phonetic recognition, we use the frame-based phone error rate as in (33) due to the additive decomposability of the Hamming distance. If the loss function is decomposed in the same manner as the joint feature map, we can add the loss function to each segment in the inference, and thus, the computational complexity for the loss-augmented inference is much reduced. Detailed explanations are given in the next subsection.

B. Stochastic Gradient Descent

It is not easy to solve the constrained optimization problem of (30) due to the large number of margin constraints: e.g. given only 40 phones, the number of possible segmentations involving 5 phonetic labels ($\ell_1\ell_2\ell_3\ell_4\ell_5$) is about 1.02×10^8. Thus, an optimization method which considers all possible number of constraints requires large computational complexity, and its implementation is difficult.

To reduce the number of constraints, optimization methods such as the soft-max approximation, cutting plane algorithm and subgradient method have been proposed [12], [13], [41], [43], [56]. In [12], [13], the large number of margin constraints associated to each training input is reduced to a single constraint by approximating the hard-max margin to the soft-max margin. In [41], [56], the cutting plane algorithm, also known as the column generation algorithm, is used to reduce the number of margin constraints by accumulating the most violating constraint in each iteration. In [43], a subgradient method which considers only the most violating constraint associated to each training input in each iteration is used. In this paper, we use two optimization methods based on the stochastic gradient descent due to its fast convergence [13], [43]: the stochastic subgradient descent using the hard-max margin and the stochastic gradient descent using the soft-max margin.

1) Stochastic Subgradient Descent using Hard-max Margin: The constrained optimization problem of (30) can be converted into an unconstrained optimization problem given by

$$\min_{\mathbf{w}} \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{w}) \tag{34}$$

where $\forall i$,

$$f_i(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\Bigg[-\langle\mathbf{w},\Phi(\mathbf{X}_i,\mathbf{y}_i)\rangle + \max_{\mathbf{y}\neq\mathbf{y}_i}\Big(\langle\mathbf{w},\Phi(\mathbf{X}_i,\mathbf{y})\rangle + \Delta(\mathbf{y}_i,\mathbf{y})\Big)\Bigg]_+ \tag{35}$$

and $[\,]_+$ denotes the hinge loss. Using the nonnegativity of the loss function in (32), the above equation can be expressed as

$$f_i(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\cdot\max_{\mathbf{y}}\Big(-\langle\mathbf{w},\Delta\Phi(\mathbf{X}_i,\mathbf{y})\rangle + \Delta(\mathbf{y}_i,\mathbf{y})\Big). \tag{36}$$

Due to the hard-max that appears in (36), $f_i(\mathbf{w})$ is not differentiable with respect to $\mathbf{w}$. Thus, we use the subgradient of $f_i(\mathbf{w})$ given by

$$\tilde{g}_i(\mathbf{w}) = \mathbf{w} - C\Delta\Phi(\mathbf{X}_i,\mathbf{y}_i^*), \quad \forall i, \tag{37}$$

where the most competing label sequence with respect to $\mathbf{w}$ is defined as

$$\mathbf{y}_i^* = \operatorname*{argmax}_{\mathbf{y}}\Big(-\langle\mathbf{w},\Delta\Phi(\mathbf{X}_i,\mathbf{y})\rangle + \Delta(\mathbf{y}_i,\mathbf{y})\Big). \tag{38}$$

Since we use a decomposable loss based on the Hamming distance in (33), a slight modification of Viterbi-like dynamic programming in (28) and (29) leads to a similar efficient inference to find $\mathbf{y}_i^*$. The stochastic subgradient descent algorithm using the hard-max margin is summarized in Algorithm 1.

Algorithm 1 Stochastic subgradient descent with hard-max

```
Choose: $\mathbf{w}_0$ and step size sequences $\{\mu_\tau\}_{\tau=1}^{\infty}$.
$\tau = 1$.
repeat
  Select a training sample $(\mathbf{X}_i,\mathbf{y}_i)$ randomly.
  Decode the most competing label sequence:
  $\mathbf{y}_i^* = \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}}\Big(-\langle\mathbf{w}_{\tau-1},\Delta\Phi(\mathbf{X}_i,\mathbf{y})\rangle + \Delta(\mathbf{y}_i,\mathbf{y})\Big)$.
  Calculate the subgradient of $f_i$:
    $\tilde{g}_i(\mathbf{w}_{\tau-1}) = \mathbf{w}_{\tau-1} - C\Delta\Phi(\mathbf{X}_i,\mathbf{y}_i^*)$.
  Update $\mathbf{w}_{\tau-1}$ by subgradient descent:
    $\mathbf{w}_\tau = \mathbf{w}_{\tau-1} - \mu_\tau\tilde{g}_i(\mathbf{w}_{\tau-1})$.
  $\tau = \tau + 1$.
until convergence
```

The exact form of the step size schedule is given as $\mu_\tau = \frac{\tau_0}{\tau+N}$, where $\tau_0 = 0.02$. This step size satisfies the Robbins-Monro conditions [57]: $\sum_{\tau=1}^{\infty}\mu_\tau = \infty$ and $\sum_{\tau=1}^{\infty}\mu_\tau^2 < \infty$. These conditions need to be satisfied for convergence.

2) Stochastic Gradient Descent using Soft-max Margin: The objective function $f_i(\mathbf{w})$ in (35) can be approximated by replacing the hard-max with the soft-max as follows:

$$f_i(\mathbf{w}) \approx \frac{1}{2}\|\mathbf{w}\|^2 + C\Big[-\langle \mathbf{w}, \Phi(\mathbf{X}_i, \mathbf{y}_i)\rangle + Z_i(\mathbf{w})\Big]_+, \forall i, \tag{39}$$

where the soft-max $Z_i(\mathbf{w})$ (a tight upper bound on the hard-max) is defined as

$$Z_i(\mathbf{w}) = \log \sum_{\mathbf{y} \neq \mathbf{y}_i} e^{(\langle \mathbf{w}, \Phi(\mathbf{X}_i, \mathbf{y})\rangle + \Delta(\mathbf{y}_i, \mathbf{y}))}. \tag{40}$$

The soft-max is differentiable with respect to $\mathbf{w}$, and the gradient of the approximated objective function is given by

$$g_i(\mathbf{w}) = \mathbf{w} - C\left[\Phi(\mathbf{X}_i, \mathbf{y}_i) - \frac{\partial Z_i(\mathbf{w})}{\partial \mathbf{w}}\right], \quad \forall i. \tag{41}$$

The gradient of the soft-max can be efficiently calculated by a dynamic programming based on the forward and backward procedures, as described in the Appendix. The stochastic gradient descent algorithm using the soft-max margin is summarized in Algorithm 2.

Algorithm 2 Stochastic gradient descent with soft-max

```
Choose: $\mathbf{w}_0$ and step size sequences $\{\mu_\tau\}_{\tau=1}^{\infty}$
$\tau = 1$
repeat
    Select a training sample $(\mathbf{X}_i, \mathbf{y}_i)$ randomly.
    Calculate the forward and backward variables.
    Calculate the gradient by (41).
    Update $\mathbf{w}_{\tau-1}$ by gradient descent:
        $\mathbf{w}_\tau = \mathbf{w}_{\tau-1} - \mu_\tau g_i(\mathbf{w}_{\tau-1})$.
    $\tau = \tau + 1$.
until convergence
```

The step size schedule for stochastic gradient descent in Algorithm 2 is same with that for stochastic subgradient descent in Algorithm 1.

IV. Experiments

We performed phonetic recognition experiments on the TIMIT speech corpus which contains 6,300 phonetically-rich utterances spoken by 630 speakers consisting of 438 males and 192 females, from 8 major dialect regions [54]. Following the standard partitioning of the corpus by National Institute of Standard Technology, we split the data into a training set (462 speakers and 3,696 utterances), development set (50 speakers and 400 utterances) and test set (118 speakers and 1,136 utterances), without overlaps [58]. The test set was again split into the traditional core test set (192 sentences) and the rest enhanced test set (944 sentences) [59]. We extracted 39-dimensional acoustic feature vectors which consist of 12 mel-frequency cepstral coefficients, log energy and the corresponding delta and acceleration coefficients, where the frame size is 25ms and the rate is 10ms. Following the standard regrouping of phonetic labels [60], 61 TIMIT phonetic labels were reduced to 48 labels, and each context-independent monophone label was represented by a one-state LMSMM, one-state LMHMM and three-state LMHMM. We initially estimated the function parameters by the ML criterion, and then we updated the estimates by large margin training based on the SSVM and the stochastic gradient descent algorithm. Note that during training, the phone boundary information was provided. Therefore, the Baum-Welch algorithm was not necessary in the initial ML training for the one-state LMSMM and one-state LMHMM. However, phonetic recognition on the development set and test set was performed by simultaneous phonetic segmentation and labeling. For the three-state LMHMM, the Baum-Welch algorithm was used in the initial ML training, and the forced alignment by the Viterbi algorithm was used for the approximated correct state-label sequence in the large margin training. The preset values, $C(> 0)$ and $R(\ell)(\in \{1, ..., 50\})$, were determined using the development set for best performance. Depending on the phonetic label, different number of bins can be used, however here we set $B(\ell) = 3, \forall \ell$ for comparisons with three-state LMHMMs.

We compare the results obtained by LMSMMs with those obtained by LMHMMs [12], [13] according to 1, 2, 4 and 8 Gaussian mixtures per bin under the same experimental setup. Note that multiple Gaussian mixtures are approximated by the single most dominant Gaussian to formulate the linear discriminant function. This is shown in (21). For the performance evaluation, 48 phonetic labels were again reduced to 39 labels [60], and then both the frame error rates based on the Hamming distances and the phone error rates based on the edit distances were calculated. Tables I and II show the frame error rates and the phone error rates on the test set, respectively, when the soft-max margin was used. For various number of mixtures, LMSMMs consistently outperformed both one-state LMHMMs and three-state LMHMMs in terms of both the frame and phone error rates. Actually, the error rates obtained by LMHMMs are slightly different from those obtained by Sha *et al.* [12], [13]. This is due to the differences in ML baselines. They also used a batch gradient descent with a line search to determine the step size in each iteration while we used a stochastic gradient descent without a line search. Recently, the LMHMM without any approximation was proposed using a variant of the bundle algorithm to solve a non-convex optimization (NCO) problem [61]. In comparison to the NCO-LMHMM [61], the performance of the LMSMM is better than that of the NCO-LMHMM. Although their bundle algorithm, which can be considered as a cutting plane algorithm, solves the original NCO problems for LMHMMs, it requires a more complex procedure involving quadratic programming, and due to the constraint accumulation, it is difficult to extend it for use in a LVCSR task.

Table III shows the phone error rates on the test set according to the hard-max margin and the soft-max margin. The LMSMMs using the soft-max margin performed better than those using the hard-max margin. Compared to LMHMMs using the hard-max margin, LMSMMs using the hard-max margin produced better results. The stochastic subgradient descent algorithm using the hard-max margin was about three times faster than the stochastic gradient descent algorithm using the soft-max margin, since the hard-max margin needs only the Viterbi recursion to find the most competing output sequence while the soft-max margin have to perform forward

TABLE I
TEST SET FRAME ERROR RATES (%) BY HAMMING DISTANCES

	Core test set				Enhanced test set			
	1-mix	2-mix	4-mix	8-mix	1-mix	2-mix	4-mix	8-mix
ML (one-state HMM)	39.5	34.4	32.5	30.8	39.2	34.0	32.1	30.5
One-state LMHMM	29.2	28.6	28.0	27.1	29.0	28.3	27.7	26.9
ML (three-state HMM)	35.4	31.1	29.4	28.6	35.3	30.9	28.9	28.3
Three-state LMHMM	29.0	27.9	27.3	26.8	28.8	27.9	27.0	26.6
ML (SMM)	32.9	29.9	28.2	27.7	32.7	29.8	28.1	27.6
LMSMM	27.9	27.2	26.8	26.4	27.5	27.1	26.7	26.3

TABLE II
TEST SET PHONE ERROR RATES (%) BY EDIT DISTANCES

	Core test set				Enhanced test set			
	1-mix	2-mix	4-mix	8-mix	1-mix	2-mix	4-mix	8-mix
ML (One-state HMM)	42.8	36.8	34.0	32.2	42.1	36.3	33.4	31.0
One-state LMHMM	31.3	30.7	29.9	28.6	30.2	29.7	29.1	28.0
ML (Three-state HMM)	37.7	33.2	30.1	29.1	37.3	32.5	29.2	28.6
Three-state LMHMM	30.2	28.8	28.0	27.6	29.5	28.2	27.6	27.2
ML (SMM)	35.9	32.1	29.6	28.5	35.1	31.3	28.9	28.1
LMSMM	28.9	28.0	27.3	27.1	28.2	27.5	27.1	26.8

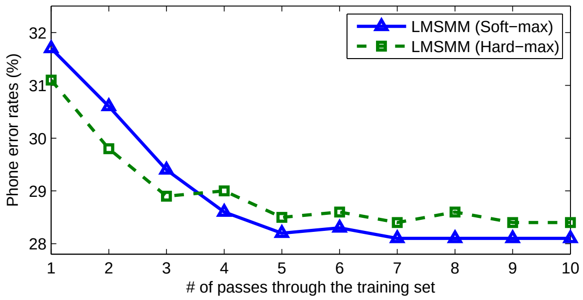

Fig. 6. Evolutions of phone error rates on the development set according to the hard-max and soft-max (LMSMM, 1-mix).

TABLE IV
PHONE ERROR RATES (%) OBTAINED BY 1-MIXTURE LMSMM ACCORDING TO DIFFERENT COMPOSITIONS OF FEATURES ON THE CORE TEST SET. NB MEANS THAT THE SEGMENT BINNING WAS NOT USED IN THE CONTENT FEATURE: $B(\ell) = 1,\ \forall \ell$.

1-mix	$\phi^d + \phi^c$	$\phi^t + \phi^c$	$\phi^t + \phi^d + \phi^c(NB)$
ML(SMM)	40.9	36.6	38.7
LMSMM (Soft-max)	31.3	29.7	30.4

TABLE V
PHONE ERROR RATES (%) OBTAINED BY PERCEPTRON TRAINING OF SMM PARAMETERS ON THE CORE TEST SET.

	1-mix	2-mix	4-mix	8-mix
SMM+Perceptron	32.5	30.7	29.1	28.4

and backward recursions and the gradient computation. However, as shown in Fig. 6, where we plot evolutions of phone error rates on the development set according to the hard-max and soft-max of 1-mixture LMSMM, the phone error rates obtained by the soft-max margin are lower than those obtained by the hard-max margin. In the hard-max margin, margin constraints for all other competing output sequences except one particular output sequence, which are the most competing with previous parameter values, are not guaranteed to be met when parameters are updated. On the other hand, the soft-max margin increases the margin between the correct output sequence and the upper bound of all competing output sequences.

Table IV shows the phone error rates obtained by 1-mixture LMSMM according to different compositions of segment features. Partial combinations achieved phone error rates higher than 28.9% obtained by the combination of whole features. Additionally, the performance of LMSMM without segment binning ($B(\ell) = 1,\ \forall \ell$) is worse than that obtained by segment binning. We also estimated the SMM parameters by the perceptron training. The performances obtained by the perceptron training are worse than those obtained by the large margin training, as shown in Table V. These comparative results show that the proposed joint feature map and the enhancement of margins scaled by Hamming loss lead great improvements in performances.

Note that the general structure, the discriminant function and the inference algorithm of the SMM are different from those of the HMM. The inference algorithm of the SMM in (28) and (29) considers both partial segmentations and segment-labelings while the HMM inference in (4) takes into account just partial frame-labelings. Therefore, even though the proposed SMM with three bins is based on similar Gaussian modeling of the observations, it produces different recognition results compared to the three-state Gaussian HMM. Moreover, the SMM framework allows averaging of the Gaussian sufficient statistics within each bin such that the SMM is less sensitive to variation in acoustic features. This averaging is in accordance with the segmental features proposed in [45], [50]. Disregarding large margin training and the proposed duration

TABLE III
TEST SET PHONE ERROR RATES (%) ACCORDING TO HARD-MAX AND SOFT-MAX

	Core test set				Enhanced test set			
	1-mix	2-mix	4-mix	8-mix	1-mix	2-mix	4-mix	8-mix
One-state LMHMM (Hard-max)	33.0	32.4	30.8	29.6	31.9	31.4	30.1	29.0
Three-state LMHMM (Hard-max)	30.8	29.8	29.0	28.5	30.2	29.3	28.7	28.1
LMSMM (Hard-max)	29.9	29.2	28.6	28.3	29.0	28.4	28.1	27.9
LMSMM (Soft-max)	28.9	28.0	27.3	27.1	28.2	27.5	27.1	26.8

feature, we experimentally show that the proposed SMM with three bins and the three-state HMM are different models leading to different performance even when both models are using similar Gaussian modeling of the observations. The ML baseline of the SMM with three bins achieved phone error rate of 36.6% (in Table IV) which is lower than 37.7% (in Table II) obtained by the ML baseline of the three-state Gaussian HMM.

By including large margin training, we notice that the performance difference between the LMSMM without duration feature and three-state LMHMM has been reduced. This suggests that large margin training had a more positive impact on the HMM than the SMM. The incorporation of the duration feature certainly improved the performance of the LMSMM but it is not clear how explicit phone duration features can be incorporated in the LMHMM framework such that the discriminant function is in linear form (a requirement for large margin training based on the SSVM). In conclusion, the performance improvement attained by the proposed LMSMM over the LMHMM is mostly attributed to the benefit of the general structure of SMM over that of HMM.

In the preliminary version [39], performance evaluations of LMSMMs were conducted only on the core test set by the hard-max margin. However, here, we used both the hard-max margin and the soft-max margin and obtained better performances on both the core test set and the enhanced set by the soft-max margin. Moreover, we also performed three-state LMHMMs for performance comparisons with LMSMMs while in the preliminary version, it was shown that LMSMMs performed better than the one-state LMHMMs.

Even though none of the LMSMMs in the experiment gives the lowest phone error rate of 23% on the core test set in the task of TIMIT phonetic recognition by complicated deep belief networks reported in [62] and the performance improvements of LMSMMs over LMHMMs become smaller as the number of mixtures increases, the proposed LMSMM is significant in that this is the first large margin discriminative model under the SMM framework for phonetic recognition that significantly improves the performance over the generative SMM. While the performances of generative SMMs are lower than those of LMHMMs, the proposed LMSMMs give better results than those obtained by LMHMMs under the same experimental setup. In addition, in comparison to the previous long-range segmental features such as the TRAP and MS features, the proposed long-range segmental feature leads to a linear discriminant function with small additional computational complexity. The linear discriminant function allows a large margin training based on the SSVM.

TABLE VI
PHONE ERROR RATES (%) OBTAINED BY BATCH LEARNING OF LMSMM PARAMETERS ON THE CORE TEST SET.

	1-mix	2-mix	4-mix	8-mix
LMSMM+Batch (Hard-max)	30.1	29.6	28.8	28.6
LMSMM+Batch (Soft-max)	29.3	28.6	27.9	27.8

Compared to the batch learning, the online learning is known to converge faster and produces a system with better generalization capability. As shown in Fig. 6, the proposed algorithm converged within 5 passes through the training set. The benefit of batch learning is that it can be performed in parallel which is important for LVCSR tasks. In the TIMIT phonetic recognition task, we performed batch learning under the proposed LMSMM framework by accumulating gradients/subgradients through the training set before updating the parameter vector. As shown in Table VI, the phone error rate of the batch learning is a little higher than that of the online learning, but it is lower than that of the three-state LMHMM.

The LMSMM has the potential to further improve its performance, since the LMSMM offers more flexibility to facilitate the incorporation of different segment-based feature maps and segmentation loss functions. The use of boundary frame features, variance features across frames and a loss as a function of segmentation boundaries might improve the performance. Furthermore, a context-dependent triphone model and a multi-state model might also improve the performance. To apply context-dependent triphone model for phonetic recognition using the proposed LMSMM framework, we need to convert monophone-based labeling to triphone-based labeling and construct a decision tree to cluster the triphones. We leave this work for the future.

A multi-state LMSMM is much more complex than the proposed one-state LMSMM with mulitple bins, since there are many possible state sequences to consider for a given phone boundary. In addition, it will be very difficult to formulate a multi-state LMSMM with a discriminant function that is in linear form. As an alternative, we consider subphone models.

Since the sub-segmentation information such as the boundaries of beginning, middle and ending segments of each phone is necessary during training, and no existing database provides this type of segmentation information, we obtained boundary segmentation information (beginning, middle and ending of each phone) using the Viterbi algorithm on a three-state LMHMM and then built a subphone LMSMM without binning. As shown in Table VII, the performance

TABLE VII
PHONE ERROR RATES (%) OBTAINED BY SUBPHONE LMSMM WITHOUT BINNING ON THE CORE TEST SET.

	1-mix	2-mix	4-mix	8-mix
subphone LMSMM (Hard-max)	29.6	29.0	28.5	28.2
subphone LMSMM (Soft-max)	28.7	27.7	27.2	26.9

is a little better than that obtained by one-state monophone LMSMM with three bins. This can be attributed to the fact that subphone LMSMM considers variable length subphones during inference. The analysis using more bins and multi-state models are left for future research.

An implemented code of the LMSMM is available at http://mmp.kaist.ac.kr/∼swkim.

V. CONCLUSION

In this paper, we propose the LMSMM for phonetic recognition. The SMM framework can be better suited for this task than the HMM framework in that SMM framework is capable of simultaneous phonetic segmentation and labeling with segment-based features. We define not a posterior probability but an explicit discriminant function and estimate the function parameters by SSVM which is a large margin learning framework for structured prediction. The proposed discriminant function is linear in the segment-based joint feature map which consists of the transition feature function, duration feature function and content feature function. As the function parameters are estimated, the SSVM increases the score margin obtained from the discriminant function by scaling it with a loss for better generalization. The stochastic gradient descent with both the hard-max margin and the soft-max margin is used to solve the optimization problem of SSVM in the primal domain due to its fast convergence and capability to handle a large number of margin constraints. Experimental results showed that the proposed LMSMM outperformed the LMHMM from experiments on the TIMIT phonetic recognition.

APPENDIX
FORWARD AND BACKWARD PROCEDURES FOR COMPUTING THE GRADIENT OF THE SOFT-MAX

The forward variable $\alpha(t,\ell)$ and the backward variable $\beta(t,\ell)$ for the i-th training sample are defined as

$$\alpha_i(t,\ell) = \sum_{\mathbf{y}' \in \mathbf{y}^{\alpha}_{t:\ell}} e^{\left(\langle \mathbf{w}, \Phi(\mathbf{X}_i,\mathbf{y}')\rangle + \Delta(\mathbf{y}_i,\mathbf{y}')\right)} \tag{42}$$

and

$$\beta_i(t,\ell) = \sum_{\mathbf{y}' \in \mathbf{y}^{\beta}_{t:\ell}} e^{\left(\langle \mathbf{w}, \Phi(\mathbf{X}_i,\mathbf{y}')\rangle + \Delta(\mathbf{y}_i,\mathbf{y}')\right)} \tag{43}$$

where $\mathbf{y}^{\alpha}_{t:\ell}$ and $\mathbf{y}^{\beta}_{t:\ell}$ denote respectively all possible partial segmentations from 1 to t such that the last segment ends at the t-th frame with label ℓ and all possible partial segmentations from $t+1$ to T such that phone ℓ transits to a certain phone at time t. The forward and backward variables are calculated recursively from the previous variables as

$$\begin{aligned}\alpha_i(t,\ell) &= \sum_{d=1}^{R(\ell)} \sum_{\ell'} \Big[\alpha_i(t-d,\ell') \\ &\qquad \times e^{\left(\langle \mathbf{w}, \phi(\ell',\ell,t-d,t,\mathbf{X}_i)\rangle + \Delta(\mathbf{y}_i,\ell,t-d,t)\right)}\Big],\end{aligned} \tag{44}$$

and

$$\begin{aligned}\beta_i(t,\ell) &= \sum_{d=1}^{R(\ell')} \sum_{\ell'} \Big[\beta_i(t+d,\ell') \\ &\qquad \times e^{\left(\langle \mathbf{w}, \phi(\ell,\ell',t,t+d,\mathbf{X}_i)\rangle + \Delta(\mathbf{y}_i,\ell',t,t+d)\right)}\Big],\end{aligned} \tag{45}$$

where the Hamming distance within a segment, which is labeled ℓ in the interval $[t_1+1, t_2]$, is given by

$$\Delta(\mathbf{y}_i,\ell,t_1,t_2)) = \sum_{t=t_1+1}^{t_2} \delta(\ell^i_t \neq \ell). \tag{46}$$

Using the forward or backward variables, we can compute the soft-max over all possible $\mathbf{y}$s including $\mathbf{y}_i$ as

$$\begin{aligned}\widetilde{Z}_i &= \log \sum_{\mathbf{y}} e^{(\langle \mathbf{w}, \Phi(\mathbf{X}_i,\mathbf{y})\rangle + \Delta(\mathbf{y}_i,\mathbf{y}))} \\ &= \log \sum_{\ell \in \mathcal{L}} \alpha_i(T,\ell) = \log \beta_i(0, start).\end{aligned} \tag{47}$$

The gradient of Z_i with respect to the m-th element of $\mathbf{w}$, w_m, is expressed as

$$\begin{aligned}\frac{\partial Z_i}{\partial w_m} &= \frac{1}{e^{\widetilde{Z}_i} - e^{\langle \mathbf{w}, \Phi(\mathbf{X}_i,\mathbf{y}_i)\rangle}} \frac{\partial\Big(e^{\widetilde{Z}_i} - e^{\langle \mathbf{w}, \Phi(\mathbf{X}_i,\mathbf{y}_i)\rangle}\Big)}{\partial w_m} \\ &= \frac{e^{\widetilde{Z}_i} \frac{\partial \widetilde{Z}_i}{\partial w_m} - e^{\langle \mathbf{w}, \Phi(\mathbf{X}_i,\mathbf{y}_i)\rangle} \phi_m(\mathbf{X}_i,\mathbf{y}_i)}{e^{\widetilde{Z}_i} - e^{\langle \mathbf{w}, \Phi(\mathbf{X}_i,\mathbf{y}_i)\rangle}}\end{aligned} \tag{48}$$

where ϕ_m is the m-th element of Φ, and

$$\begin{aligned}\frac{\partial \widetilde{Z}_i}{\partial w_m} &= \frac{1}{e^{\widetilde{Z}_i}} \sum_{t=1}^{T} \sum_{\ell \in \mathcal{L}} \sum_{d=1}^{R(\ell)} \sum_{\ell'} \phi_m(\ell',\ell,t-d,t,\mathbf{X}_i)\alpha_i(t-d,\ell') \\ &\quad \times \beta_i(t,\ell) e^{\left(\langle \mathbf{w}, \phi(\ell',\ell,t-d,t,\mathbf{X}_i)\rangle + \Delta(\mathbf{y}_i,\ell,t-d,t)\right)}.\end{aligned} \tag{49}$$

ACKNOWLEDGMENT

The authors would like to thank Dr. Alex Smola for valuable discussions on LMSMM.

REFERENCES

[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.

[2] B.-H. Juang, W. Chou, and C. H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 3, pp. 257–265, 1997.

[3] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. IEEE ICASSP*, 2002.

[4] H. Jiang, X. Li, and C. Liu, "Large margin hidden markov models," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1584–1595, 2006.

[5] J. Li, M. Yuan, and C. H. Lee, "Approximate test risk bound minimization through soft margin estimation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2393–2404, 2007.

[6] X. Li and H. Jiang, "Solving large-margin hidden markov model estimation via semidefinite programming," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2283–2392, 2007.

[7] D. Yu, L. Deng, X. He, and A. Acero, "Use of incrementally regulated discriminative margins in mce training for speech recognition," in *Interspeech*, 2006.

[8] ——, "Large-margin minimum classification error training for large-scale speech recognition tasks," in *Proc. IEEE ICASSP*, 2007.

[9] H. Jiang and X. Li, "Incorporating training errors for large margin hmms under semi-definite programming framework," in *Proc. IEEE ICASSP*, 2007.

[10] V. Vapnik, *The nature of statistical learning theory*. Springer, 2000.

[11] J. Li, Z.-J. Yan, C.-H. Lee, and R.-H. Wang, "A study on soft margin estimation for lvcsr," in *Proc. IEEE ASRU*, 2007.

[12] F. Sha and L. K. Saul, "Large margin hidden markov models for automatic speech recognition," in *NIPS*, 2007.

[13] F. Sha, "Large margin training of acoustic models for speech recognition," *Ph.D. thesis, Univ. Pennsylvania*, 2007.

[14] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random fields for phone classification," in *Interspeech*, 2005.

[15] Y.-H. Sung and D. Jurafsky, "Hidden conditional random fields for phone recognition," in *Proc. IEEE ASRU*, 2009.

[16] J. Morris and E. Fosler-Lussier, "Conditional random fields for integrating local discriminative classifiers," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 617–628, 2008.

[17] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001.

[18] M. Ostnedorf, V. Digalakis, and O. Kimball, "From hmms to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 5, pp. 360–378, 1996.

[19] S.-Z. Yu, "Hidden semi-markov models," *Artificial Intelligence*, vol. 174, pp. 215–243, 2010.

[20] S. E. Levinson, "Continuously variable duration hidden markov models for automatic speech recognition," *Computer Speech and Language*, vol. 1, pp. 29–45, 1986.

[21] M. Johnson, "Capacity and complexity of hmm duration modeling techniques," *IEEE Signal Processing Letters*, vol. 12, no. 5, pp. 407–410, 2005.

[22] J. Pylkkönen and M. Kurimo, "Duration modeling techniques for continuous speech recognition," in *Interspeech*, 2004.

[23] S. Roucos, M. Ostendorf, H. Gish, and A. Derr, "Stochastic segment modeling using the estimate-maximize algorithm," in *Proc. IEEE ICASSP*, 1988.

[24] M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 12, pp. 1857–1869, 1989.

[25] H. Gish and K. Ng, "A segmental speech model with applications to word spotting," in *Proc. IEEE ICASSP*, 1993.

[26] M. Russell and W. Holmes, "Linear trajectory segmental hmms," *IEEE Signal Processing Letters*, vol. 4, no. 3, pp. 72–74, 1997.

[27] R. Chengalvarayan, "Linear trajectory models incorporating preprocessing parameters for speech recognition," *IEEE Signal Processing Letters*, vol. 5, no. 1, pp. 66–68, 1998.

[28] L. Deng, M. Aksmanovic, D. Sun, and J. Wu, "Speech recognition using hidden markov models with polynomial regression functions as nonstationary states," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 507–520, 1994.

[29] M. Russell and P. Jackson, "A multiple-level linear/linear segmental hmm with a formant-based intermediate layer," *Computer Speech and Language*, vol. 19, pp. 205–225, 2005.

[30] M. Gales and S. Young, "Segmental hmm's for speech recognition," in *Proc. Euro. Conf. Speech Commun. Technol.*, 1993.

[31] L. Deng, D. yu, and A. Acero, "Structured speech modeling," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1492–1504, 2006.

[32] C.-F. Li, M.-H. Siu, and J. S.-K. Au-Yeung, "Recursive likelihood evaluation and fast search algorithm for polynomial segment model with application to speech recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1704–1718, 2006.

[33] V. Digalakis and M. Ostendorf, "Fast algorithms for phone classification and recognition using segment-based models," *IEEE Trans. Speech and Audio Processing*, vol. 40, no. 12, pp. 2885–2896, 1992.

[34] W. Goldenthal, "Statistical trajectory models for phonetic recognition," *Ph.D. thesis, M.I.T.*, 1994.

[35] J. Frankel, "Linear dynamic models for automatic speech recognition," *Ph.D. thesis, University of Edinburgh*, 2003.

[36] Q. Shi, L. Wang, L. Cheng, and A. Smola, "Discriminative human action segmentation and recognition using semi-markov model," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[37] O. Thomas, P. Sunehag, G. Dror, S. Yun, S. Kim, M. Robards, A. Smola, D. Green, and P. Saunders, "Wearable sensor activity analysis using semi-markov models with a grammar," *Pervasive and Mobile Computing*, vol. 6, pp. 342–350, 2010.

[38] S. Sarawagi and W. W. Cohen, "Semi-markov conditional random fields for information extraction," in *NIPS*, 2005.

[39] S. Kim, S. Yun, and C. Yoo, "Large margin training of semi-markov model for phonetic recognition," in *Proc. IEEE ICASSP*, 2010.

[40] G. Zweig and P. Nguyen, "Scarf: A segmental crf speech recognition system," in *Technical Report*, 2009.

[41] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and independent output variables," *Journal of Machine Learning Research 6*, 2005.

[42] F. Sha and L. K. Saul, "Comparison of large margin training to other discriminative methods for phonetic recognition by hidden Markov models," in *Proc. IEEE ICASSP*, 2007.

[43] N. Ratliff, J. A. Bagnell, and M. Zinkevich, "(online) subgradient methods for structured prediction," in *AISTATS*, 2007.

[44] J. Goldberger, D. Burshtein, and H. Franco, "Segmental modeling using a continuous mixture of nonparametric models," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 3, pp. 262–271, 1999.

[45] J. R. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, vol. 17, pp. 137–152, 2003.

[46] W. Holmes and M. Russell, "Probabilistic-trajectory segmental hmms," *Computer Speech and Language*, pp. 3–37, 1999.

[47] D. Burshtein, "Robust parametric modeling of durations in hidden markov models," in *Proc. IEEE ICASSP*, 1995.

[48] G. Heigold, R. Schlüter, and H. Ney, "On the equivalence of gaussian hmm and gaussian hmm-like hidden conditional random fields," in *Interspeech*, 2007.

[49] M. Layton, "Augmented statistical models for classifying sequence data," *Ph.D. thesis, Univ. Cambridge*, 2006.

[50] L. Tóth, "Posterior-based speech models and their application to hungarian speech recognition," *Ph.D. thesis, Univ. Szeged*, 2007.

[51] H. Hermansky and S. Sharma, "Traps: classifiers of temporal patterns," in *Proc. ICSLP*, 1998.

[52] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, pp. 117–132, 1998.

[53] V. Tyagi, I. McCowan, H. Bourlard, and H. Misra, "Mel-cepstrum modulation spectrum (mcms) features for robust asr," in *Proc. IEEE ASRU*, 2003.

[54] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa timit acoustic phonetic continuous speech corpus cdrom," in *NIST*, 1993.

[55] B. Taskar, C. Guestrin, and D. Koller, "Max-margin markov networks," in *NIPS*, 2004.

[56] T. Joachims, T. Finley, and C. N. J. Yu, "Cutting-plane training of structural svms," *Machine Learning*, pp. 1–33, 2009.

[57] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, pp. 400–407, 1951.

[58] A. K. Hallberstadt and J. R. Glass, "Heterogeneous acoustic measurements for phonetic clasification," in *Eurospeech*, 1997.

[59] I. Heintz, E. Fosler-Lussier, and C. Brew, "Discriminative input stream combination for conditional random field phone recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1533–1546, 2009.

[60] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. Acoustic, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[61] T.-M.-T. Do and T. Artières, "Large margin training for hidden markov models with partially observed states," in *ICML*, 2009.

[62] A. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009.

Sungwoong Kim (S'07) received the B.S. degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2004. He is currently pursuing his Ph. D. degree in the Department of Electrical Engineering, KAIST. His research interest is machine learning for signal processing.

Sungrack Yun (S'06) received the B.S. degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2003. He is currently pursuing his Ph. D. degree in the Department of Electrical Engineering, KAIST. His research interest is machine learning for signal processing.

Chang D. Yoo (S'92-M'96) received the B.S. degree in Engineering and Applied Science from California Institute of Technology in 1986, the M.S. degree in Electrical Engineering from Cornell University in 1988 and the Ph.D. degree in Electrical Engineering from Massachusetts Institute of Technology in 1996. From January 1997 to March 1999 he worked at Korea Telecom as a Senior Researcher. He joined the Department of Electrical Engineering at Korea Advanced Institute of Science and Technology in April 1999. From March 2005 to March 2006, he was with Research Laboratory of Electronics at MIT. His current research interests are in the application of machine learning and digital signal processing in multimedia. He is a member of Tau Beta Pi and Sigma Xi. Prof. Yoo currently serves on the Machine Learning for Signal Processing (MLSP) Technical Committee of the IEEE Signal Processing Society.