LARGE MARGIN TRAINING OF SEMI-MARKOV MODEL FOR PHONETIC RECOGNITION

Sungwoong Kim, Sungrack Yun, and Chang D. Yoo

Department of Electrical Engineering, Korea Advanced Institute of Science & Technology

sungwoong.kim01@gmail.com, yunsungrack@kaist.ac.kr, cdyoo@ee.kaist.ac.kr

ABSTRACT

This paper considers a large margin training of semi-Markov model (SMM) for phonetic recognition. The SMM framework is better suited for phonetic recognition than the hidden Markov model (HMM) framework in that the SMM framework is capable of simultaneously segmenting the uttered speech into phones and labeling the segment-based features. In this paper, the SMM framework is used to define a discriminant function that is linear in the joint feature map which attempts to capture the long-range statistical dependencies within a segment and between adjacent segments of variable length. The parameters of the discriminant function are estimated by a large margin learning criterion for structured prediction. The parameter estimation problem, which is an optimization problem with many margin constraints, is solved by using a stochastic subgradient descent algorithm. The proposed large margin SMM outperforms the large margin HMM on the TIMIT corpus.

Index Terms— Hidden Markov model, semi-Markov model, structured support vector machine, phonetic recognition.

1. INTRODUCTION

Over the past two decades, a continuous-density hidden Markov model, which is considered a probabilistic generative model, has been extensively used for automatic speech recognition (ASR). A generative hidden Markov model (HMM) for ASR represents the joint probability of the observation (acoustic feature vector extracted from one frame) sequence and label (word or phone) sequence under the assumptions that the frame-based labels follow a Markov process and adjacent frame-based features are conditionally independent given the corresponding label. Therefore, generative HMMs are restricted such that they cannot capture long-range statistical dependencies across frames, and the HMM parameters estimated by maximizing the joint probability does not lead to minimum prediction error rate. This has led to interest in the following two HMMs: i) the discriminatively-trained generative HMMs [1–3] which adopt discriminative training algorithms to train generative HMMs; ii) the discriminative HMMs [4–6] which define a non-probabilistic discriminant function or directly represent the posterior probability under the HMM frameworks. Although the discriminatively-trained generative HMMs and the discriminative HMMs have been shown to yield better prediction accuracies than generatively-trained generative HMMs, these are limited to modeling local statistical dependencies between adjacent observations and predicting a label for each observation without explicit segmentation under the HMM frameworks.

The semi-Markov model (SMM) framework, on the other hand, is based on a segment-based Markovian structure, and this framework is capable of simultaneously segmenting and labeling sequential data. The SMM can capture long-range statistical dependencies within a segment and between adjacent segments of variable length. Thus, the SMM framework is considered a more appropriate framework compared to the HMM framework for ASR, since a joint segmentation and labeling is required in ASR.

Previously, SMM frameworks for ASR have been limited to incorporating an explicit duration model into the generative HMM, leading to a generative SMM [7–9]. This moderate extension has not received full benefits of the SMM capability to capture segment-based rich features. In addition, a generative training of model parameters does not attempt to minimize the prediction error rate on unseen data. Thus, the performance improvements of the previously proposed generative SMMs over the generative HMMs are lower than the improvements obtained by the discriminative HMMs over the generative HMMs.

In this paper, we propose a large margin SMM (LMSMM) for phonetic recognition. In the task of phonetic recognition, a sequence of phonetic labels should be predicted from a speech utterance without any given segmentation information. We simultaneously perform phonetic segmentation and labeling with segment-based features under the SMM framework. The proposed LMSMM is in contrast to the semi-Markov CRFs [10, 11] in that we define not a posterior probability but an explicit discriminant function and estimate the function parameters by structured support vector machine (SSVM) [12] which is a large margin learning criterion for structured prediction and is considered to have better generalization ability than other learning criteria for structured prediction [13]. The proposed discriminant function is linear in the segment-based joint feature map which is composed 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. The parameter estimation problem leads to an optimization problem with many margin constraints. The stochastic subgradient descent [14] is used to solve the optimization problem of SSVM in the primal domain, since it guarantees fast convergence and it can handle a large number of margin constraints. The TIMIT phonetic recognition task is performed to evaluate the proposed LMSMM. Experimental results show that the proposed LMSMM outperforms the large margin HMM (LMHMM) [4]. In addition, comparative results show that the proposed joint feature map and the large margin training lead to better performances than other joint feature maps and the perceptron training, respectively.

The rest of this paper is organized as follows. Section 2 describes the proposed discriminative SMM for phonetic recognition.

This work was in part supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Culture Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2009, and was in part supported (National Robotics Research Center for Robot Intelligence Technology, KAIST) by Ministry of Knowledge Economy under Human Resources Development Program for Convergence Robot Specialists.

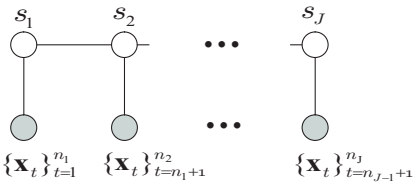

Fig. 1. An undirected graph of discriminative SMM.

Section 3 explains the large margin training based on the SSVM and the stochastic subgradient descent algorithm. Section 4 evaluates the proposed LMSMM with a number of experimental and comparative results. Section 5 concludes the paper.

2. DISCRIMINATIVE SEMI-MARKOV MODEL FOR PHONETIC RECOGNITION

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 an utterance having a length of T frames, such that

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

where $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ 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.[1] Throughout this paper, F is assumed to be linear in $\mathbf{w}$, and it is mathematically represented as

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

where $\Phi(\mathbf{X}, \mathbf{y}) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^M$ is the joint feature map which maps a paired input and output sequence into an M-dimensional feature space.

As shown in Fig. 1, a discriminative SMM assumes a segment-based Markovian structure and predicts a sequence of phonetic labels with explicit phonetic segmentation. A variable number of frames are assigned to one hidden state representing a phonetic segment, and the behavior within a segment is non-Markovian. Therefore, $\mathbf{y}$ can be 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, $\ell_j(\in \mathcal{L})$, and J denote the ending frame of the j-th segment (such that $n_{j+1} > n_j, \forall j$, and $n_J = T$), the phonetic label of the j-th segment, and the number of segments, respectively. Note that the number of segments J itself is a variable. In a discriminative SMM, s_j is dependent only on s_{j-1}, s_{j+1}, and the acoustic feature vectors in the j-th segment $\{\mathbf{x}_t\}_{t=n_{j-1}+1}^{n_j}$. This segment-based Markovian property decomposes a joint feature map $\Phi(\mathbf{X}, \mathbf{y})$ into a sum over segment features ϕ as

$$\Phi(\mathbf{X}, \mathbf{y}) = \sum_{j=1}^{J} \phi(\ell_{j-1}, \ell_j, n_{j-1}, n_j, \{\mathbf{x}_t\}_{t=n_{j-1}+1}^{n_j}). \tag{3}$$

2.1. Segment features

Segment features characterize the statistical dependencies 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

$$\begin{aligned}\phi(\ell_{j-1}, \ell_j, n_{j-1}, n_j, \{\mathbf{x}_t\}_{t=n_{j-1}+1}^{n_j}) \\ = [(\phi^t(\ell_{j-1}, \ell_j))^T, (\phi^d(\ell_j, n_{j-1}, n_j))^T, \\ (\phi^c(\ell_j, n_{j-1}, n_j, \{\mathbf{x}_t\}_{t=n_{j-1}+1}^{n_j}))^T\,]^T.\end{aligned} \tag{4}$$

The transition feature for phonetic transition from ℓ' to ℓ, $\phi^t_{(\ell',\ell)}$ is defined as an indicator function to capture the statistical dependencies between two neighboring phonetic segments under the SMM framework:

$$\phi^t_{(\ell',\ell)}(\ell_{j-1}, \ell_j) = \delta(\ell_{j-1} = \ell',\ \ell_j = \ell) \tag{5}$$

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.

The duration feature for phone ℓ, ϕ^d_ℓ is defined as the sufficient statistics for the gamma distribution, since the gamma distribution is considered good for modeling the phone duration [7, 8]:

$$\phi^d_\ell(\ell_j, n_{j-1}, n_j) = \begin{bmatrix} \log(n_j - n_{j-1}) \\ n_j - n_{j-1} \\ 1 \end{bmatrix} \delta(\ell_j = \ell). \tag{6}$$

The content feature is defined by both the labeled segment and all observations within a phone segment. The discriminative SMM allows a non-Markovian behavior within a segment to construct a segment-based content feature that captures long-range statistical dependencies on inputs. We divide a segment into a number of bins and 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 define $\mathrm{vec}(A)$ as the $((D+1)(D+2)/2)$-dimensional vector of upper triangular elements from A. The content feature for phone ℓ and the k-th bin, $\phi^c_{(\ell,k)}$, is given by

$$\begin{aligned}\phi^c_{(\ell,k)}(\ell_j, n_{j-1}, n_j, \{\mathbf{x}_t\}_{t=n_{j-1}+1}^{n_j}) \\ = \frac{B(\ell)}{n_j - n_{j-1}} \sum_{t\in b_k} \mathrm{vec}\left(\begin{bmatrix} \mathbf{x}_t\mathbf{x}_t^T & \mathbf{x}_t \\ \mathbf{x}_t^T & 1 \end{bmatrix} \right) \delta(\ell_j = \ell)\end{aligned} \tag{7}$$

where

$$\begin{aligned} b_k \;=\; & \{n_{j-1} + \frac{n_j - n_{j-1}}{B(\ell)}(k-1) + 1, ..., \\ & n_{j-1} + \frac{n_j - n_{j-1}}{B(\ell)} k\}, \quad k \in \{1, ..., B(\ell)\}, \end{aligned} \tag{8}$$

and $B(\ell)$ denotes the number of bins according to the phonetic label ℓ. Since, the statistical characteristics of acoustic feature vectors may vary within a phonetic segment, we use binning features by dividing a segment into a number of bins and modeling each bin differently. Moreover, we take averages within a bin to become insensitive to small changes of acoustic feature vectors.[2]

2.2. Efficient 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 $t - d$. The recursion of

[1] Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{L}$ be the space of the acoustic feature vector sequence, phonetic label sequence, and phonetic label, respectively.

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

the Viterbi-like dynamic programming for efficient SMM inference is given by

$$U(t,\ell) = \underset{(d,\ell')\in\{1,\ldots,R(\ell)\}\times\mathcal{L}}{\operatorname{argmax}} \Big(V(t-d,\ell')+ \langle \mathbf{w}, \phi(\ell',\ell,t-d,t,\{\mathbf{x}_u\}_{u=t-d+1}^{t})\rangle\Big), \quad (9)$$

$$V(t,\ell) = \max_{(d,\ell')\in\{1,\ldots,R(\ell)\}\times\mathcal{L}} \Big(V(t-d,\ell')+ \langle \mathbf{w}, \phi(\ell',\ell,t-d,t,\{\mathbf{x}_u\}_{u=t-d+1}^{t})\rangle\Big) \quad (10)$$

where $R(\ell)$ is the restricted range of admissible durations of phone ℓ for 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 Eq. (10) requires $O(T|\mathcal{L}|\sum_\ell R(\ell))$ computations of $\langle \mathbf{w}, \phi\rangle$, and this computational cost is not much higher in comparison to $O(|\mathcal{L}|^2 T)$ computations for the HMM inference.

3. LARGE MARGIN TRAINING

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 Eq. (1) leads to the minimum prediction error rate on unseen data. In this paper, we use a large margin learning criterion for structured prediction, SSVM [12], and adopt the stochastic subgradient descent [14] to solve the optimization problem of SSVM.

3.1. Structured support vector machine

The SSVM estimates $\mathbf{w}$ by minimizing a quadratic objective function subject to a set of linear soft margin constraints as follows [12]:

$$\min_{\mathbf{w},\xi} \ \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{i=1}^{N}\xi_i \quad (11)$$

$$\text{s.t.} \quad \langle \mathbf{w}, \Phi(\mathbf{X}_i,\mathbf{y}_i)\rangle - \max_{\mathbf{y}\neq\mathbf{y}_i}\langle \mathbf{w}, \Phi(\mathbf{X}_i,\mathbf{y})\rangle \geq \Delta(\mathbf{y}_i,\mathbf{y}) - \xi_i$$

$$\xi_i \geq 0, \ \ \forall i,$$

where $C > 0$ 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. Even though the string-based phone error rate by edit distance is a more proper measure for phonetic recognition, we use the Hamming distance based on frame errors as a loss function due to its decomposability for a loss-augmented inference in a stochastic subgradient descent algorithm.

3.2. Stochastic subgradient descent

It is not easy to solve the constrained optimization problem of (11) due to the large number of margin constraints and the hard-max term in margin constraints. In this paper, we use the stochastic subgradient descent [14] due to its fast convergence and robustness in handling a large number margin constraints and the hard-max.

The constrained optimization problem of (11) can be converted into an unconstrained optimization problem:

$$\min_{\mathbf{w}} \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{w}) \quad (12)$$

where $f_i(\mathbf{w})$ is given by

$$f_i(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\Bigg[-\langle \mathbf{w}, \Phi(\mathbf{X}_i,\mathbf{y}_i)\rangle + \max_{\mathbf{y}\neq\mathbf{y}_i}\Big(\langle \mathbf{w}, \Phi(\mathbf{X}_i,\mathbf{y})\rangle + \Delta(\mathbf{y}_i,\mathbf{y})\Big)\Bigg]_+, \quad (13)$$

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

$$f_i(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\cdot\max_{\mathbf{y}}\Big(-\langle \mathbf{w}, \Delta\Phi(\mathbf{X}_i,\mathbf{y})\rangle + \Delta(\mathbf{y}_i,\mathbf{y})\Big) \quad (14)$$

where $\Delta\Phi(\mathbf{x}_i,\mathbf{y}) = \Phi(\mathbf{x}_i,\mathbf{y}_i) - \Phi(\mathbf{x}_i,\mathbf{y})$. We solve this unconstrained optimization problem by stochastic subgradient descent algorithm which is described as follows:

Algorithm: Stochastic subgradient descent

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

4. EXPERIMENTS

We performed phonetic recognition experiments on the TIMIT speech corpus, which was divided into the training set (462 speakers and 3,696 utterances), development set (50 speakers and 400 utterances), and core test set (24 speakers and 192 utterances), without overlaps. 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 grouping of phonetic labels, 61 phonetic labels were reduced to 48 labels, and each label was represented by one state in the discriminative SMM. Initially, we estimated the function parameters by the maximum likelihood (ML) criterion, and then we updated those by the large margin training. The preset values, $C(>0)$, $R(\ell)(\in\{1,...,50\})$, and $B(\ell)(\in\{1,...,5\})$ were determined using the development set for the best performance. When evaluating the performance on the core test set, we used parameters $\mathbf{w}$ that achieved the best result on the development set, reduced 48 phonetic labels to 39 labels, and calculated the phone error rates based on the edit distances. The performance comparisons between LMSMMs and LMHMMs [4] were carried out using different number of Gaussian mixtures under the same experimental setup. [3]

[3]Note that multiple Gaussian mixtures are approximated by the single most dominant Gaussian to formulate the linear discriminant function.

Table 1. Phone error rates (%) on the core test set.

	ML (HMM)	LMHMM	ML (SMM)	LMSMM
1-mix	42.8	34.2	37.4	30.0
2-mix	36.8	32.6	32.7	28.7
4-mix	34.3	30.7	30.8	28.5
8-mix	32.1	29.9	29.5	28.5

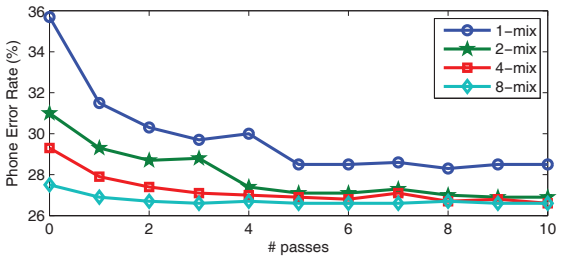

Fig. 2. The evolution of the phone error rates of LMSMM on the development set.

Tab. 1 shows the phone error rates on the core test set. LMSMMs consistently outperformed LMHMMs across all model sizes. Note that Fig. 2 shows the evolution of the phone error rates obtained by LMSMM on the development set. We observed the fast convergence of the stochastic subgradient descent algorithm: most improvements occur within 5 passes through the training set. In Tab. 2, we give the phone error rates obtained by 1-mixture LMSMM according to different compositions of segment features. The combination of the transition and content feature shows a better performance than the combination of the duration and content feature. But, both achieved results higher than 30.0% obtained by LMSMM with the combination of whole features. Additionally, the advantage of the segment binning in the content feature is verified: the performance of LMSMM without segment binning ($B(\ell) = 1,\ \forall \ell$) was worse than that obtained by segment binning.

Furthermore, we estimated the SMM parameters by the perceptron training, which is equal to setting a loss $\Delta(\mathbf{y}_i, \mathbf{y})$ to zero for all $\mathbf{y}$s in the large margin training. As shown in Tab. 3, the performances of the perceptron training was worse than those of the large margin training. This demonstrates that the enhancement of margins scaled by Hamming loss leads great improvements in performances.

5. CONCLUSION

In this paper, we proposed the LMSMM for phonetic recognition. Under the SMM framework, we defined a linear discriminant function and segment-based joint feature map which consists of the transition feature, duration feature, and content feature. The function parameters were estimated by the large margin training based on the SSVM and the stochastic subgradient descent algorithm. Experimental results showed that the proposed LMSMM outperformed the LMHMM on the TIMIT phonetic recognition.

Table 2. Phone error rates (%) of 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)	41.9	36.8	39.7
LMSMM	32.3	30.7	31.4

Table 3. 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	34.3	31.8	29.9	29.1

6. REFERENCES

[1] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *ICASSP*, 2002, vol. 1, pp. 105–108.

[2] 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.

[3] 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.

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

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

[6] 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.

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

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

[9] K. Oura, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "A fully consistent hidden semi-markov model-based speech recognition system," *IEICE Trans. Information and Systems*, vol. E91-D, no. 11, pp. 2693–2700, 2008.

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

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

[12] 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.

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

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