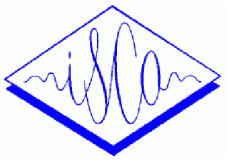


INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Efficient Segmental Conditional Random Fields for Phone Recognition

Yanzhang He, Eric Fosler-Lussier

Department of Computer Science and Engineering
The Ohio State University, Columbus, OH, USA
{hey, fosler}@cse.ohio-state.edu

Abstract

Recently the initial attempt has been made to use segment-based direct models on their own for phone classification and recognition without the aid of an HMM lattice. This paper follows this line of research to further investigate these one-pass segmental direct models on phone recognition using posteriors as input. We make the first direct comparison between a frame-based system and a segmental system using the same base features, and explore the utilization of transition features in a direct segmental model for the first time. The results show that transition features can be very beneficial, particularly the ones surrounding the segment boundaries. In order to efficiently incorporate such features, we propose the Boundary-Factored SCRF, which reduces the time complexity of a Segmental Conditional Random Field (SCRF) to that of a frame-level CRF.

Index Terms: Segmental Conditional Random Fields, Phone Recognition

10.21437/Interspeech.2012-515

1. Introduction

Discriminative segmental models have been a promising recent direction of speech recognition research [1, 2, 3]. Typically these models have improved on state-of-the-art systems by relying on a first-pass lattice from an HMM system to significantly constrain candidate segmentations and label sequences. Only recently have direct segment-based models been explored as an initial pass for phone classification and recognition [4]. We further investigate such direct utilization of Segmental CRFs in continuous feature spaces for phone recognition.

Frame-based systems have been used for one-pass phone recognition (e.g. [5]), but as far as we are aware there has been no direct comparison, using the same basic feature set, of conditional models that are either frame-based or segmental in nature. We explore the same feature set as in [5] within a segment-based system, and find that segmental features constructed simply through sub-sampling frame posteriors or taking average/maximum/minimum frame posteriors over the entire segment show significant improvement over frame-based system. Indicators of segment durations are also experimented as segment-based bias features.

Discriminative sequence models such as CRFs have the ability to have state transitions depend on observations; however these features are seldom exploited in segmental models. Zweig has found that features surrounding segment boundaries are particularly useful for phone classification and recognition but he only uses them as segment state features [4]. Morris has shown the effectiveness of transition features but on a frame-based system [6]. This paper made initial attempts to study the effect of the observation-dependent transition features in segmental models. The results show that transition features are very important for segmental models, especially the local ones surrounding segment boundaries independent of the whole segment, which are in effect better than segmental transition features.

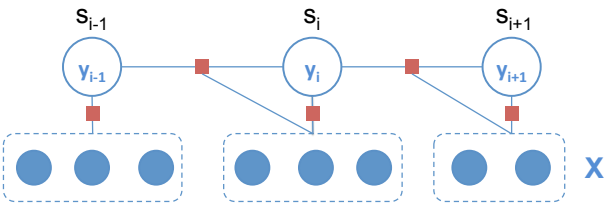
However, it is expensive to use transition features directly in SCRFs. We further propose Boundary-Factored SCRF (BF-SCRF), an efficient SCRF model, for the transition features only surrounding segment boundaries or even no transition features (only transition bias). The proposed model reduces the training and decoding time by a factor of maximum duration, leading to a SCRF with the same level of time complexity as a frame-level CRF.

2. Models

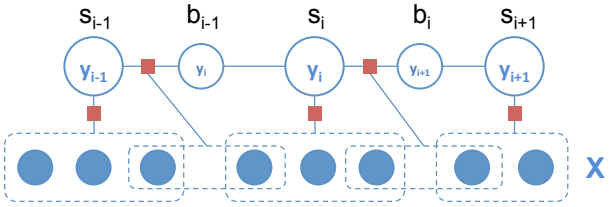
2.1. Segmental CRFs

Segmental Conditional Random Fields, also known as SCRFs [2] or semi-Markov CRFs [7], are sequence models in which each state can span a variable non-unit length of time. As is shown in Figure 1(a), the label states at the top in a SCRF correspond to segments of observations at the bottom with different time span which can capture high-level features. The Markov assumption is only held on the transitions between segment states, while the transitions within segments can be non-Markovian.

The SCRF model structure is defined by segmentation: all T frames in the observation sequence $\mathbf{x}$ may be segmented into $p \leqslant T$ chunks in all possible ways. Let $\mathbf{s} = \langle s_1, s_2, \cdots, s_p \rangle$ denote a sequence of p segments with time and label information; $\mathbf{e} = \langle e_1, e_2, \cdots, e_p \rangle$ denote a time sequence, where e_j is the ending time of s_j; $\mathbf{y} = \langle y_1, y_2, \cdots, y_p \rangle$ denote a segment-level label se-



(a) SCRFs. Transition features could be segmental.



(b) BF-SCRFs. Transition features base on fixed boundary windows.

Figure 1: *SCRFs and Boundary-Factored SCRFs Graphical Models.*

quence, where y_j is the label for s_j. Define each segment $s_j = \langle y_j, e_{j-1}, e_j \rangle$, where $1 \leqslant j \leqslant p, e_0 = 0$. Adjacent segments touch and their lengths are always positive, so $0 \leqslant e_{j-1} < e_j \leqslant T, \forall j\ 1 \leqslant j \leqslant p$.

Then the joint probability distribution of the label sequence and its associated segmentation conditioned on the observations is modeled as

$$P(\mathbf{y}, \mathbf{e}|\mathbf{x}) = \frac{\exp \sum_{j=1}^{|\mathbf{e}|} \sum_i \lambda_i f_i(y_{j-1}, y_j, x_{e_{j-1}}^{e_j})}{Z(\mathbf{x})} \quad (1)$$

where $Z(\mathbf{x})$ is the normalization term. $x_{e_{j-1}}^{e_j}$ denotes the segment of $\mathbf{x}$ starting from frame e_{j-1} (exclusive) to e_j (inclusive). f_i is a function defined on a pair of adjacent segment labels and the corresponding observed segment.

Like linear-chain CRFs, parameters for SCRFs are estimated by optimizing the Conditional Maximum Likelihood (CML), i.e. $P(\mathbf{y}, \mathbf{e}|\mathbf{x})$. The gradient of the log-likelihood is:

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L} = \mathbf{F}(\mathbf{y}, \mathbf{e}, \mathbf{x}) - \sum_{\substack{\mathbf{y}', \mathbf{e}'\ s.t. \\ |\mathbf{e}|=|\mathbf{y}|}} P(\mathbf{y}', \mathbf{e}'|\mathbf{x}) \mathbf{F}(\mathbf{y}', \mathbf{e}', \mathbf{x}) \quad (2)$$

One has a choice whether to optimize only the label sequence $P(\mathbf{y}|\mathbf{x})$ (as in [4]), or the joint label and segmentation space $P(\mathbf{y}, \mathbf{e}|\mathbf{x})$ (as in [7]). In the former case, one needs to sum over all possible segmentations $\mathbf{e}'$ for $\mathbf{y}$ leading to a non-convex likelihood; it is an open question whether the marginalization of segmentations is necessary for most tasks. We choose to jointly model $\mathbf{y}$ and $\mathbf{e}$, requiring that the correct (or force-aligned) segmentation $\mathbf{e}$ is also known during training, but assures a global optimum using gradient-descent based optimization; an extended forward-backward algorithm permits efficient parameter estimation and inference, described below.

Let $\alpha_e^{Seg}(y)$ be defined as the accumulated potentials from the beginning of the sequence up to time e for all the paths that end with a segment ending at time e and labeled y. Let $\beta_e^{Seg}(y)$ be defined as the backward-accumulated potentials from the end of the sequence back to time $(e+1)$ for all the paths that start with a segment ending at time e and labeled y. Then the original alpha-beta recursion would be modified as:

$$\alpha_e^{Seg}(y) = \sum_{d=1}^{D} \sum_{y'} \alpha_{e-d}^{Seg}(y') \boldsymbol{\lambda}^\top \mathbf{f}(y', y, x_{e-d}^{e}) \quad 1 \leq e \leq T \quad (3)$$

$$\beta_e^{Seg}(y) = \sum_{d=1}^{D} \sum_{y'} \boldsymbol{\lambda}^\top \mathbf{f}(y, y', x_{e}^{e+d}) \beta_{e+d}^{Seg}(y') \quad 1 \leq e < T \quad (4)$$

where $\alpha_0^{Seg}(y) = 1$, $\beta_T^{Seg}(y) = 1$, $\mathbf{f}$ is the vector of all feature functions, and D is the pre-defined upper bound of segment duration. As in linear-chain CRFs, we can then efficiently compute the gradient and the normalization term $Z(\mathbf{x})$ with the given alphas and betas (see [7]).

Replacing summations with maximizations in the alpha recurrence yields the forward algorithm for decoding.

According to (3) and (4), the time complexity of training and decoding for SCRFs is $O(TN^2DM)$, where N is the size of the label space $\mathcal{Y}$, M is the total number of feature functions. Note that this is slower than frame-level CRFs with a factor of D, whose time complexity is $O(TN^2M)$.

2.2. Boundary-Factored Segmental CRFs

Due to the time complexity of SCRFs, using transition features directly can be expensive. We propose Boundary-Factored SCRFs (BF-SCRFs), (Figure 1(b)), which introduce a redundant boundary state node in between every pair of adjacent segment states in a SCRF. The boundary state carries the exact label of its succeeding segment, but only spans a fixed-size window of frames centered at the last frame of its preceding segment state. A formal description of BF-SCRFs is given below:

Let $\mathbf{s}$, $\mathbf{e}$ and $\mathbf{y}$ denote the same as previous section. In addition, $\mathbf{b} = \langle b_1, b_2, \cdots, b_{p-1} \rangle$ denotes a boundary sequence of $\mathbf{x}$, where $b_j = \langle y_{j+1}, e_j \rangle$ is the boundary (with fixed-size context) between segments s_j and s_{j+1}. Then a segmentation with label assignment of $\mathbf{x}$ is denoted by $\mathbf{r} = \langle s_1, b_1, s_2, b_2, \cdots, s_{p-1}, b_{p-1}, s_p \rangle$. Let $g_i(y_{j-1}, y_j, x_{e_{j-1}-c}^{e_{j-1}+c})$ denote the feature function between s_{j-1} and b_{j-1}, where c is the left/right context size of the boundary window; $h_k(y_j, x_{e_{j-1}}^{e_j})$ denote the feature function between b_{j-1} and s_j. Each g_i and h_k is associated with a weight μ_i and ν_k respectively. The prob-

ability of a segment sequence $\mathbf{r} = (\mathbf{y}, \mathbf{e})$ conditioned on the observed data sequence $\mathbf{x}$ is thus calculated as:

$$P(\mathbf{y}, \mathbf{e}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\{\sum_{j=1}^{|\mathbf{e}|}(\sum_i \mu_i g_i(y_{j-1}, y_j, x_{e_{j-1}-c}^{e_{j-1}+c}) + \sum_k \nu_k h_k(y_j, x_{e_{j-1}}^{e_j}))\} \quad (5)$$

Training for BF-SCRFs based on CML estimation is done by forward-backward algorithm as well. Let $\alpha_e^{Seg}(y)$ be the alpha for a segment state, the same as in a general SCRF; $\alpha_{e',e}^{Seg}(y)$ represent accumulated potentials with all paths from the beginning of the sequence up to time e that end with a segment spanning from time e' to e and labeled y; $\alpha_e^{Bound}(y)$ be the alpha for a boundary state, representing the accumulated potentials with all paths from the beginning of the sequence up to time e that end with a segment ending at time e and transitioning into a label y. Alpha recurrence is thus derived as:

$$\alpha_e^{Seg}(y) = \sum_{d=1}^{D} \alpha_{e-d,e}^{Seg}(y) = \sum_{d=1}^{D} \alpha_{e-d}^{Bound}(y) \boldsymbol{\nu}^\top \mathbf{h}(y, x_{e-d}^{e}) \quad (6)$$

$$\alpha_e^{Bound}(y) = \sum_{y'} \alpha_e^{Seg}(y') \boldsymbol{\mu}^\top \mathbf{g}(y', y, x_{e_{j-1}-c}^{e_{j-1}+c}) \quad (7)$$

The beta recurrence can be derived similarly. For decoding, we just need to replace all the summations in eq(6) and (7) with maximizations. If M is the maximum number of g_i and h_k, then the complexity of training and decoding for BF-SCRFs is $O(T(ND + N^2)M)$, which is $O(2TN^2M)$ assuming $D < N$. This represents an $O(D)$ speedup over general SCRFs. When transition feature functions used in a SCRF are only local features (allowing a fixed-size context window) or even no transition features (only transition bias), BF-SCRFs can be used to speed up SCRF implementations.

3. Features

In order to directly compare SCRFs and frame CRFs, we construct the segment-level features by simple utilization of frame-level features across the segment in various ways described below. Two kinds of base acoustics input are used: (1) MLP phone and/or phonological posteriors $\text{PP}_\theta(x_t)$ (as in [5]), where θ takes all possible phone and/or phonological classes, e.g. PHONE=/b/ or MANNER=nasal; (2) MLP phone and/or phonological boundary posteriors $\text{BP}_\theta(x_t)$ (as in [8]), where θ could be PHONE_BOUND=yes or VOICE_BOUND=no, etc.

Segmental state features are functions of a segment label and the observed segment data. All features of this kind (except the duration feature) are in the form of

$$h_{y,\theta}(y_j, x_{e_{j-1}}^{e_j}) = \phi_\theta(x)\delta(y_j = y) \quad (8)$$

- Sample Feature: $\phi_{\theta,\Delta t}(x) = \text{PP}_\theta(x_{e_{j-1}+\Delta t})$, $\Delta t = \{0.1l, 0.3l, 0.5l, 0.7l, 0.9l\}$, $l = e_j - e_{j-1}$.
- Avg Feature: $\phi_\theta(x) = (\sum_{i=e_{j-1}+1}^{e_j} \text{PP}_\theta(x_i))/l$.
- Max Feature: $\phi_\theta(x) = \max_{i=e_{j-1}+1}^{e_j} \text{PP}_\theta(x_i)$.
- Min Feature: $\phi_\theta(x) = \min_{i=e_{j-1}+1}^{e_j} \text{PP}_\theta(x_i)$.
- Duration Feature:
  $h_{y,d}(y_j, x_{e_{j-1}}^{e_j}) = \delta(y_j = y)\delta(e_j - e_{j-1} = d)$, for all (y, d), where $y \in \mathcal{Y}$ and $1 \leqslant d \leqslant D$.

Segmental transition features are functions of a segment of the observed data, and a pair of consecutive segment labels. They are in the form of:

$$f_{y',y,\theta}(y_{j-1}, y_j, x_{e_{j-1}}^{e_j}) = \phi_\theta(x)\delta(y_{j-1} = y')\delta(y_j = y) \quad (9)$$

$\phi_\theta(x)$ for segmental transition features used in our experiments are exactly the same as for segmental state features defined above.

Boundary transition features are functions of the observed data surrounding the boundary within a fixed-size context window, the left and the right labels to the boundary. They are in the form of

$$g_{y',y,\theta,\Delta t}(y_{j-1}, y_j, x_{e_{j-1}-c}^{e_{j-1}+c}) = \phi_{\theta,\Delta t}(x)\delta(y_{j-1} = y')\delta(y_j = y) \quad (10)$$

where $-c < \Delta t \leqslant c$, and $\phi_{\theta,\Delta t}(x) = \text{PP}_\theta(x_{e_{j-1}} + \Delta t)$ or $\text{BP}_\theta(x_{e_{j-1}} + \Delta t)$.

4. Experiments

We evaluate frame-level CRFs, standard SCRFs and BF-SCRFs using the TIMIT phone recognition task: we trained MLPs on phone and phonological classes from TIMIT phonetic transcriptions (61 phone labels, 44 phonological attributes) as in [5]. Each CRF system is trained on the posterior functions using stochastic gradient decent; regularization is done by weight averaging across iterations. All systems are monophone-based. Following standard practice we build models for 48 phones and map to 39 phones for scoring purposes. We use a partitioning of the test set into development set (50 speaker), core set (24 speaker), and enhanced set (118 speaker including core set) described in [9].

The baseline frame-level CRF [5] directly uses frame-level phone and/or phonological posteriors. The maximum duration $D = 10$ was chosen according to SCRF performance on the development set. For only *segmental state features*, only the transition bias is used but not transition features. The experiment results show that segmental CRFs achieve significant performance gain over

Table 1: *Phone recognition accuracy for different models based on MLP phone posteriors only / phone+phonological posteriors. Significant ($p<0.05$) improvement at $\sim$0.9% for development, $\sim$1.4% for core and $\sim$0.6% for enhanced set.*

Model	Features	Dev Accuracy %	Core Accuracy %	Enhanced Accuracy %	Train time per epoch with only phone posteriors
frame CRFs	frame state ftr (1)	71.2 / 72.3	69.0 / 70.2	70.0 / 71.4	5m
frame CRFs	(1) + frame transition ftr	71.6 / 72.6	69.8 / 70.8	70.5 / 71.7	11m
BF-SCRFs	segmental state ftr (2)	73.1 / 73.3	71.0 / 71.3	71.9 / 72.2	16m (SCRFs: 62m) [1]
SCRFs	(2) + segmental transition ftr	74.4 / 74.9	72.3 / 72.7	73.0 / 73.6	580m
BF-SCRFs	(2) + boundary transition ftr (c=1)	74.3 / 75.0	72.1 / 72.5	73.1 / 73.3	33m (SCRFs: 235m) [1]
BF-SCRFs	(2) + boundary transition ftr (c=6)	75.2 / 76.0	73.5 / 73.5	73.9 / 74.4	108m (SCRFs: 915m) [1]

frame-level CRFs. As in [5], for segmental CRFs, the combination of the phone posteriors and phonological posteriors works better than phone posteriors alone.

Only *boundary transition features* can enable the faster implementation of SCRFs (BF-SCRFs), while the standard SCRFs must be used for *segmental transition features*. Both types of transition features help to improve the phone recognition accuracy, but boundary transition features with larger window size are better while remaining more efficient than segmental transition features. The best feature combination for our segmental CRFs achieves 73.5% phone accuracy on the core test set, which suggests that this style of feature representation may provide some advantage over detector-oriented features explored in [4] (where 66.9% core recognition accuracy was reported).

Training times in Table 1 for the phone posterior systems are measured on a single thread 3.40GHz Xeon E31270 CPU with 16G memory. Comparing models with only state features, the BF-SCRF uses only 3 times the training time as frame level system per training epoch, despite having eight times as many features. When considering transition features, BF-SCRFs (with context size 1) provide comparable performance to SCRFs (with segmental transition features) with an order of magnitude speedup. For all kinds of features where the BF-SCRF can be used, it runs faster than SCRF with up to 8 times. The amount of speedup is similar for decoding, although the exact decoding time is not reported due to space limit.

In the future, we plan to move to full word decoding; in a preliminary study we performed phone recognition on the Wall Street Journal 5k task. Space does not permit full detail of the experimental procedure (see [6] for details), but we found that BF-SCRFs achieved a 78.4% phone accuracy on WSJ0, compared to 71.0% and 74.0% using 1-state and 3-state frame-level CRFs respectively, suggesting that we may be able to improve the word-level decoder in [6], as well providing good first-pass decoders for general SCRF systems [2].

[1] SCRFs have the same accuracies but run much slower.

5. Conclusion

We introduce Boundary-Factored SCRFs, which allow efficient use of transition features in segmental CRFs. Using this model, we demonstrated the effectiveness of transition features, and provided a direct comparison with frame-level CRFs using the same feature set. The proposed model reduces the time complexity of a SCRF by a factor proportional to maximum segment duration, leading to the same computational complexity as a frame-level CRF, if transition features are local features independent of the entire segment. In the future, we plan to extend our promising phone recognition results into word recognition.

Acknowledgements: This work was supported by NSF Grant IIS-0643901 (CAREER). The authors would like to thank Jeremy Morris for providing frame CRFs code.

6. References

[1] M. Layton and M. Gales, "Augmented statistical models for speech recognition," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I.

[2] G. Zweig and P. Nguyen, "A segmental CRF approach to large vocabulary continuous speech recognition," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 152–157.

[3] S. Zhang, A. Ragni, and M. Gales, "Structured log linear models for noise robust speech recognition," *Signal Processing Letters, IEEE*, vol. 17, no. 11, pp. 945–948, 2010.

[4] G. Zweig, "Classification and recognition with direct segment models," in *Proc. ICASSP*, Kyoto, Japan, 2012.

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

[6] J. Morris, "A study on the use of conditional random fields for automatic speech recognition," Ph.D. dissertation, The Ohio State University, 2010.

[7] S. Sarawagi and W. Cohen, "Semi-markov conditional random fields for information extraction," *Advances in Neural Information Processing Systems*, vol. 17, pp. 1185–1192, 2004.

[8] Y. Wang and E. Fosler-Lussier, "Integrating phonetic boundary discrimination explicitly into hmm systems," in *Ninth International Conference on Spoken Language Processing*, 2006.

[9] A. Halberstadt and J. Glass, "Heterogeneous acoustic measurements for phonetic classification," in *Fifth European Conference on Speech Communication and Technology*, 1997.