Investigations into the Crandem Approach to Word Recognition

Rohit Prabhavalkar, Preethi Jyothi, William Hartmann, Jeremy Morris, and Eric Fosler-Lussier
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH
{prabhava,jyothi,hartmanw,morrijer,fosler}@cse.ohio-state.edu

Abstract

We suggest improvements to a previously proposed framework for integrating Conditional Random Fields and Hidden Markov Models, dubbed a Crandem system (2009). The previous authors' work suggested that local label posteriors derived from the CRF were too low-entropy for use in word-level automatic speech recognition. As an alternative to the log posterior representation used in their system, we explore frame-level representations derived from the CRF feature functions. We also describe a weight normalization transformation that leads to increased entropy of the CRF posteriors. We report significant gains over the previous Crandem system on the Wall Street Journal word recognition task.

1 Introduction

Conditional Random Fields (CRFs) (Lafferty et al., 2001) have recently emerged as a promising new paradigm in the domain of Automatic Speech Recognition (ASR). Unlike Hidden Markov Models (HMMs), CRFs are direct discriminative models: they predict the probability of a label sequence conditioned on the input. As a result, CRFs can capture long-range dependencies in the data and avoid the need for restrictive independence assumptions. Variants of CRFs have been successfully used in phone recognition tasks (Gunawardana et al., 2005; Morris and Fosler-Lussier, 2008; Hifny and Renals, 2009).

While the improvements in the phone recognition task are encouraging, recent efforts have been directed towards extending the CRF paradigm to the word recognition level (Zweig and Nguyen, 2009; Morris and Fosler-Lussier, 2009). The Crandem system (Morris and Fosler-Lussier, 2009) is one of the promising approaches in this regard. The Crandem system is directly inspired by the techniques of the Tandem system (Hermansky et al., 2000), where phone-label posterior estimates produced by a Multi-Layer Perceptron (MLP) are transformed into a suitable acoustic representation for a standard HMM. In both systems, the frame-based log posterior vector of $P(\text{phone}|\text{acoustics})$ over all phones is decorrelated using the Karhunen-Loeve (KL) transform; unlike MLPs, CRFs take into account the entire label sequence when computing local posteriors. However, posterior estimates from the CRF tend to be overconfident compared to MLP posteriors (Morris and Fosler-Lussier, 2009).

In this paper, we analyze the interplay between the various steps involved in the Crandem process. Is the local posterior representation from the CRF the best representation? Given that the CRF posterior estimates can be overconfident, what transformations to the posteriors are appropriate?

In Section 2 we briefly describe CRFs and the Crandem framework. We suggest techniques for improving Crandem word recognition performance in Section 3. Details of experiments and our results are discussed in Sections 4 and 5 respectively. We conclude with a discussion of future work in Section 6.

2 CRFs and the Crandem System

Conditional random fields (Lafferty et al., 2001) express the probability of a label sequence Q conditioned on the input data X as a log-linear sum of

weighted feature functions,

$$p(Q|X) = \frac{\exp \sum_t \sum_j \lambda_j s_j(q_t, X) + \sum_j \mu_j f_j(q_{t-1}, q_t, X)}{Z(X)} \quad (1)$$

where $s_j(\cdot)$ and $f_j(\cdot)$ are known as state feature functions and transition feature functions respectively, and λ_j and μ_j are the associated weights. $Z(X)$ is a normalization term that ensures a valid probability distribution. Given a set of labeled examples, the CRF is trained to maximize the conditional log-likelihood of the training set. The log-likelihood is concave over the entire parameter space, and can be maximized using standard convex optimization techniques (Lafferty et al., 2001; Sha and Pereira, 2003). The local posterior probability of a particular label can be computed via a forward-backward style algorithm. Mathematically,

$$p(q_t = q|X) = \frac{\alpha_t(q|X)\beta_t(q|X)}{Z(X)} \quad (2)$$

where $\alpha_t(q|X)$ and $\beta_t(q|X)$ accumulate contributions associated with possible assignments of labels before and after the current time-step t. The Crandem system utilizes these local posterior values from the CRF analogously to the way in which MLP-posteriors are treated in the Tandem framework (Hermansky et al., 2000), by applying a log transformation to the posteriors. These transformed outputs are then decorrelated using a KL-transform and then dimensionality-reduced to be used as a replacement for MFCCs in a HMM system. While the MLP is usually reduced to 39 dimensions, the standard CRF benefits from a higher dimensionality reduction (to 19 dimensions). The decorrelated outputs are then used as an input representation for a conventional HMM system.

3 Improving Crandem Recognition Results

Morris and Fosler-Lussier (2009) indicate that the local posterior outputs from the CRF model produces features that are more heavily skewed to the dominant phone class than the MLP system, leading to an increase in word recognition errors. In order to correct for this, we perform a non-linear transformation on the local CRF posterior representation before applying a KL-transform and subsequent stages. Specifically, we normalize all of the weights λ_j and μ_j in Equation 1 by a fixed positive constant n to obtain normalized weights λ'_j and μ'_j. We note that the probability of a label sequence computed using the transformed weights, $p'(Q|X)$, is equivalent to taking the nth-root of the CRF probability computed using the unnormalized weights, with a new normalization term $Z'(X)$

$$p'(Q|X) = \frac{p(Q|X)^{1/n}}{Z'(X)} \quad (3)$$

where, $p(Q|X)$ is as defined in Equation 1. Also observe that the monotonicity of the nth-root function ensures that if $p(Q_1|X) > p(Q_2|X)$ then $p'(Q_1|X) > p'(Q_2|X)$. In other words, the rank order of the n-best phone recognition results are not impacted by this change. The transformation does, however, increase the entropy between the dominant class from the CRF and its competitors, since $p'(Q|X) < p(Q|X)$. As we shall discuss in Section 5, this transformation helps improve word recognition performance in the Crandem framework.

Our second set of experiments are based on the following observation regarding the CRF posteriors. As can be seen from Equation 2, the CRF posteriors involve a global normalization over the entire utterance as opposed to the local normalization of the MLP posteriors in the output softmax layer. This motivates the use of representations derived from the CRF that are 'local' in some sense. We therefore propose two alternative representations that are modeled along the lines of the linear outputs from an MLP. The first uses the sum of the state feature functions, to obtain a vector $f^{\text{state}}(X, t)$ for each time step t and input utterance X of length $|\mathcal{Q}|$ dimensions, where $\mathcal{Q}$ is the set of possible phone labels

$$f^{\text{state}}(X, t) = \left[\sum_j \lambda_j s_j(q, X) \right]^T \qquad \forall q \in \mathcal{Q} \quad (4)$$

where q is a particular phone label. Note that the lack of an exponential term in this representation ensures that the representation is less 'spiky' than the CRF posteriors. Additionally, the decoupling of the representation from the transition feature functions could potentially allow the system to represent rel-

ative ambiguity between multiple phones hypothesized for a given frame.

The second 'local' representation that we experimented with incorporates the CRF transition feature functions as follows. For each utterance X we perform a Viterbi decoding of the most likely state sequence $Q^{\text{best}} = \text{argmax}_Q\{p(Q|X)\}$ hypothesized for the utterance X. We then augmented the state feature representation with the sum of the transition features corresponding to the phone label hypothesized for the previous frame (q_{t-1}^{best}) to obtain a vector $f^{\text{trans}}(X,t)$ of length $|\mathcal{Q}|$,

$$f^{\text{trans}}(X,t) = \left[\sum_j \lambda_j s_j(q,X) + \sum_j \mu_j f_j(q_{t-1}^{\text{best}}, q, X)\right]^T \quad (5)$$

As a final note, following (Morris and Fosler-Lussier, 2009), our CRF systems are trained using the linear outputs of MLPs as its state feature functions and transition biases as the transition feature functions. Hence, f^{state} is a linear transformation of the MLP linear outputs down to $|\mathcal{Q}|$ dimensions.[1] Both f^{state} and f^{trans} can thus be viewed as an implicit mapping performed by the CRF of the input feature function dimensions down to $|\mathcal{Q}|$ dimensions. Note that the CRF implicitly uses information concerning the underlying phone labels unlike dimensionality reduction using KL-transform.

4 Experimental Setup

To evaluate our proposed techniques, we carried out word recognition experiments on the speaker-independent portion of the Wall Street Journal 5K closed vocabulary task (WSJ0). Since the corpus is not phonetically transcribed, we first trained a standard HMM recognition system using PLP features and produced phonetic transcriptions by force aligning the training data. These were used to train an MLP phone classifier with a softmax output layer, using a 9-frame window of PLPs with 4000 hidden layer units to predict one of the 41 phone labels (including silence and short pause). The linear outputs of the MLP were used to train a baseline Tandem system. We then trained a CRF using the MLP linear outputs as its state feature functions. We extract local posteriors as well as the two 'local' representations described in Section 3. These input representations were then normalized at the utterance level, before applying a KL-transformation to decorrelate them and reduce dimensionality to 39 dimensions. Finally, each of these representations was used to train a HMM system with intra-word triphones and 16 Gaussians per mixture using the Hidden Markov Model Toolkit (Young et al., 2002).

[1] We note that our system uses an additional state bias feature that has a fixed value of 1. However, since this is a constant term, it has no role to play in the derived representation.

System	Accuracy (%)
Crandem-baseline	89.4%
Tandem-baseline	91.8%
Crandem-NormMax	91.4%
Crandem-Norm5	92.1%
Crandem-state	91.7%
Crandem-trans	91.0%

Table 1: Word recognition results on the WSJ0 task

5 Results

Results for each of the experiments described in Section 4 are reported in Table 1 on the 330-sentence standard 5K non-verbalized test set. The Crandem-baseline represents the system of (Morris and Fosler-Lussier, 2009). Normalizing the CRF weights of the system by either the weight with largest absolute value (CRF-NormMax) or by 5 (tuned on the development set) leads to significant improvements ($p \leq 0.005$) over the Crandem baseline. Similarly, using either the state feature sum (Crandem-state) or the representation augmented with the transition features (Crandem-trans) leads to significant improvements ($p \leq 0.005$) over the Crandem baseline. Note that the performance of these systems is comparable to the Tandem baseline.

To further analyze the results obtained using the state feature sum representations and the Tandem baseline, we compute the mean distance for each phone HMM from every other phone HMM obtained at the end of the GMM-HMM training phase. The distance between two HMMs is computed as a uniformly weighted sum of the average distances between the GMMs of a one-to-one alignment of states corresponding to the two HMMs. GMM distances are computed using a 0.5-weighted sum of inter-dispersions normalized by self-dispersions (Wang et

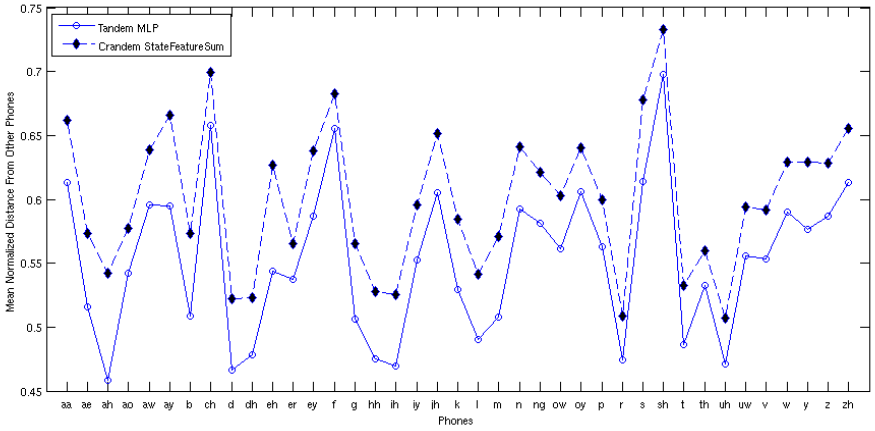

Figure 1: Normalized mean distances for each of the phone models from every other phone model trained using the Tandem MLP baseline and the state feature sum representation.

al., 2004). Distances between monomodal Gaussian distributions were computed using the Bhattacharyya distance measure. The phone HMM distances are normalized using the maximum phone distance for each system. As can be seen in Figure 1, the mean distances obtained from the state feature sum representation are consistently greater than the corresponding distances in the Tandem-MLP system, indicating larger separability of the phones in the feature space. Similar trends were seen with the transition feature sum representation.

6 Conclusions and Future Work

In this paper, we report significant improvements over the Crandem baseline. The weight normalization experiments confirmed the hypothesis that increasing the entropy of the CRF posteriors leads to better word-level recognition. Our experiments with directly extracting frame-level representations from the CRF reinforce this conclusion. Although our experiments with the systems using the state feature sum and transition feature augmented representation did not lead to improvements over the Tandem baseline, the increased separability of the phone models trained using these representations is encouraging. In the future, we intend to examine techniques by which these representations could be used to further improve word recognition results.

Acknowledgement: *The authors gratefully acknowledge support by NSF grants IIS-0643901 and IIS-0905420 for this work.*

References

A. Gunawardana, M. Mahajan, A. Acero, and J. Platt. 2005. Hidden conditional random fields for phone classification. Interspeech.

H. Hermansky, D. Ellis, and S. Sharma. 2000. Tandem connectionist feature stream extraction for conventional hmm systems. ICASSP.

Y. Hifny and S. Renals. 2009. Speech recognition using augmented conditional random fields. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(2):354–365.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML.

J. Morris and E. Fosler-Lussier. 2008. Conditional random fields for integrating local discriminative classifiers. *IEEE Transactions on Acoustics, Speech, and Language Processing*, 16(3):617–628.

J. Morris and E. Fosler-Lussier. 2009. Crandem: Conditional random fields for word recognition. Interspeech.

F. Sha and F. Pereira. 2003. Shallow parsing with conditional random fields. NAACL.

Xu Wang, Peng Xuan, and Wang Bingxi. 2004. A gmm-based telephone channel classification for mandarin speech recognition. ICSP.

S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. 2002. *The HTK Book*. Cambridge University Press.

G. Zweig and P. Nguyen. 2009. A segmental crf approach to large vocabulary continuous speech recognition. ASRU.