CRANDEM SYSTEMS: CONDITIONAL RANDOM FIELD ACOUSTIC MODELS FOR HIDDEN MARKOV MODELS

Eric Fosler-Lussier, Jeremy Morris

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

ABSTRACT

In recent years, Conditional Random Fields (CRFs) have been examined as a statistical model for speech recognition. In this paper, we explore the use of features derived via CRFs as inputs to a Tandem-style HMM ASR system (that is, a *Crandem* system). We present a model for deriving frame-level posterior features via CRFs to use in Crandem modeling and additionally provide experimental results that show the Crandem system can slightly significantly outperform both a comparable Tandem system and a comparable CRF system on the task of phone recognition.

***Index Terms*—**

Speech recognition, Stochastic fields, Feature extraction, Hidden Markov models

1. CRFS AND THE PATH TO LARGE VOCABULARY RECOGNITION

Conditional Random Fields (CRFs)[1] have been recently making some inroads in the field of automatic speech recognition; they can be seen as an extension of Hidden Markov Models, in that the parameters of a single-Gaussian HMM can be directly represented within a particular form of CRF – a "linear chain" CRF – where first-order Markovian dependencies of state sequences are maintained [2]. There are several theoretical advantages of CRFs over HMMs that have been touted elsewhere [1, 2, 3], including a naturally discriminative training criterion, lack of independence assumptions about consecutive frames of input, and the ability to incorporate negative evidence. Because of the log-linear form of the model, it is straightforward to integrate different representations of the acoustics for the speech recognition problem, including traditional acoustic features, such as MFCCs [2], affinity scores for gaussian models [4], posteriors for phonetic classes [3] and phonological features [3, 5], or phone transition estimates [6].

Unfortunately, to this point, CRF systems have been used exclusively in the realm of phone classification or phone recognition, particularly on the TIMIT dataset. This is in part due to one of the chief disadvantages of CRFs: the most straightforward, matrix-oriented implementation requires estimation of $O(N^2)$ parameters, where N is the number of state labels. Implementing a biphone or triphone-based CRF system will require sparse matrix techniques, and possibly rethinking of the discriminative criterion that is used to train CRFs, as one may not wish to discriminate between instances of the same phone in different contexts. An active area of debate is whether the state sequence is in fact the appropriate place to model context; some (including the authors) postulate that it may be possible to model contextual influence in the input to the CRF, thus alleviating the need for increasing the state space. Furthermore, the straightforward implementation of more complex, mixture of Gaussian acoustic models requires including a hidden state in the model – feasible, but more computationally intensive [2].

While phone classification and recognition has been useful for comparing the behavior of various training algorithms and inputs, if CRFs are to make it into the mainstream there must be a path forward to large-vocabulary word recognition. There are several potential methodologies that we are exploring. Recasting the word recognition problem into a pure-CRF framework that is amenable to the type of acoustic modeling provided by phone-level CRFs will require new decoding strategies; given the relatively short time of development of CRF-based modeling compared to HMM-based modeling, it will likely be years before the CRFs can catch up (and hopefully surpass) HMM models on large vocabulary tasks.

However, one potential solution for the interim is to take inspiration from Tandem acoustic models [7], in which neural network frame-level posteriors of (e.g.) phone classes are suitably modified to serve as observations for HMM-based systems. These new acoustic features can be used alone, or more commonly, with traditional features in standard mixture-of-Gaussian HMM recognizers [8]. The system derived from training HMMs on local posteriors produced by CRFs can be called a *Crandem* system, in light of its similarity to the Tandem system.

In this paper we describe some initial experiments that compare performance on TIMIT phone recognition for different decoding strategies using PLP coefficients, Multi-Layer Perceptron (MLP) posterior estimates of phone classes, and MLP posterior estimates of phonological features. In particular, we investigate whether the CRF improvements over Tandem systems that we have previously reported [3] hold up when the CRF posteriors are used in a Crandem system. Thus, while the eventual target is word recognition, this initial study remains in the phone recognition domain in order to make the appropriate comparisons. The next section is dedicated to describing the process of extracting local posterior functions for HMMs from MLPs and CRFs; this is followed by a description of the experiments, and the results and conclusions following from these experiments.

2. DERIVING LOCAL POSTERIOR FUNCTIONS FOR HMMS

In the Tandem approach [7], the acoustic input X is transformed into a more discriminative representation of the input signal via a transformation function $X' = F(X)$ before submitting these features to an HMM system. The form of F investigated in the paper cited above uses local frame posteriors derived from an MLP as a basis for the transformation, which incorporate not only the local frame

but the surrounding context frames as acoustic evidence. The MLP estimates $P(q_i|X_{i-c}^{i+c})$, the acoustic probability of being in state q at time i given the acoustics in the surrounding $\pm c$ frames. Two particularly successful instantiations of $F(X)$ from [7] include

$$F(X) = \text{KLT}(\log P(q_i|X_{i-c}^{i+c}))$$
$$F(X) = \text{KLT}(\text{linearize}(P(q_i|X_{i-c}^{i+c})))$$

where KLT stands for a Karhunen-Loève transform, and the *linearize* operation strips away the softmax output of the MLP, leaving just the weighted linear sum of connections from the hidden layer to the output layer.

We also use this input transformation $F(X)$ in the CRF training paradigm: parameters are estimated to maximize the conditional log likelihood of the joint sequence of labels Q given some representation of the input X. The probability expression takes the form of

$$P(Q|X) = \frac{\exp(\sum_i \sum_t \lambda_i f_i(q_{t-1}, q_t, X, t))}{Z(X)}$$

where the f_i represent functions of pairs of states, the acoustic input, and time t, λ_i is a learned weight for the function, and $Z(X)$ is a normalization constant over all possible paths corresponding to the input X. In a linear chain CRF, we can separate the f_i into *state* functions s_j (with weights λ_j), which associate input with a single state, and *transition* functions t_k (with weights μ_k), which associate input with pairs of states.

$$P(Q|X) = \frac{\exp(\sum_t \sum_j \lambda_j s_j(q_t, X, t) + \sum_k \mu_k t_k(q_{t-1}, q_t, X, t))}{Z(X)}$$

In our previous work [3], we used MLP posterior estimates directly as state feature functions; the transition feature functions did not have a direct correspondence to input.[1] More specifically, if the MLP provided posteriors for 61 phone classes, we defined a state feature function that associated every posterior estimate with each CRF state label. The system learns, in part, not only the association between the true label and the posterior estimate for that label provided by the MLP, but also the confusions that are made by the MLP.

Following [9], however, one can also make the transitions dependent on the acoustic input; in some of the experiments described here we use the self-same MLP posteriors as transition functions. This allows for some small amount of context-dependence while utilizing monophone CRF labels. Clearly, one can (and should) design feature functions that are more attuned to transitions, such as phonetic and phonological feature boundaries [6].

In a manner similar to that for HMMs, the joint probability of the label sequence $P(Q|X)$ in a linear chain CRF can be computed using a forward-backward algorithm. Using the matrix notation of [10], we can define a set of transition matrices M_t where

$$M_t[q, q'] = \exp(\sum_j \lambda_j s_j(q', X, t) + \sum_k \mu_k t_k(q, q', X, t)),$$

that is, the transition matrix for a particular time t is composed of the exponentiated sum of the feature functions between every pair of

[1]However, bias features were used that provided transition information akin to the state-to-state transition probabilities in an HMM.

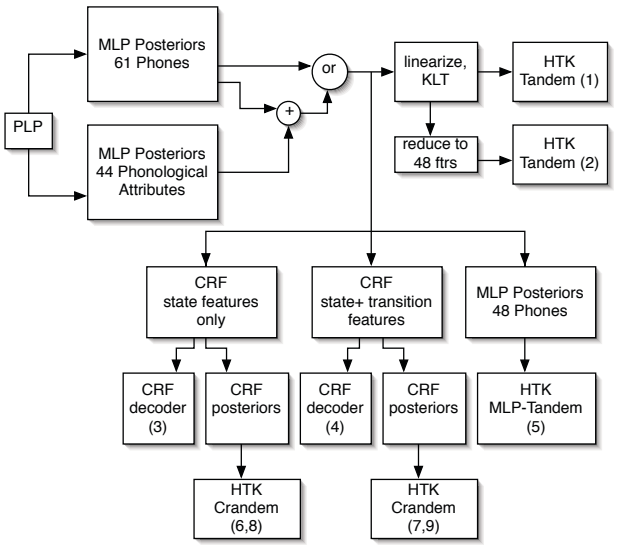

Fig. 1. Training regimens for each of the 9 systems described in Table 1. System numbers from the text and table are given in parentheses.

transitions, with appropriate terms added for the state functions. This makes it simple to define an alpha-beta recurrence for the forward-backward computation [10]:

$$\boldsymbol{\alpha}_t = \begin{cases} \boldsymbol{\alpha}_{t-1} M_t, 0 < t \leq n \\ \mathbf{1}, t = 0 \end{cases}, \boldsymbol{\beta}_{t-1}^T = \begin{cases} M_t \boldsymbol{\beta}_t^T, 1 \leq t < n \\ \mathbf{1}^T, t = n \end{cases}$$

The local posterior $P(q_{i,t}|X)$ for any label i at time t can be calculated directly from the alphas and betas [1]:

$$P(q_{i,t}|X) = \frac{\alpha_{i,t}\beta_{i,t}}{Z(X)}, Z(X) = \sum_j \alpha_{j,t}\beta_{j,t}$$

As noted above, in the Tandem framework, posteriors are never directly passed into the HMM as observation probabilities; rather, usually some form of whitened log posterior is used, either by directly taking the log of the posterior (suitably flooring for $\log(0)$) and performing a Karhunen-Loève transform, or by replacing the softmax function with a linear function, and performing a similar KLT. The first strategy may be directly replicated in the CRF paradigm – HMM systems built with this strategy are notated $\text{Crandem}_{\log}$. Another option, similar to the softmax-replacement in the Tandem system, is to ignore the normalization constant; taking $\log \alpha_{i,t}\beta_{i,t}$ in the CRF system is similar in spirit to utilizing linear MLP outputs in place of softmax outputs. This is referred to as $\text{Crandem}_{\text{unnorm}}$.

3. EXPERIMENTAL DESIGN

The primary focus of our experiments is to see if utilizing CRF posteriors within an HMM framework will continue to provide the same level of performance as the standard CRF framework especially across diverse input representations. However, there were a number of empirical questions that occurred during our experimental design phase, so we trained nine different forms of TIMIT phone recognizers which had a common input. These systems are outlined in Figure 1.

	System	Dev	Core	Ext
	PLP HMM reference	69.7	67.4	68.1
1	Tandem (61 ftrs)	72.1	69.4	70.6
2	Tandem (48 ftrs)	72.6	69.6	70.8
3	CRF (state only)	71.1	68.9	69.9
4	CRF (state+trans)	71.4	69.5	70.7
5	MLP-Tandem	70.0	67.2	68.2
6	Crandem$_{\log}$ (state)	72.9	69.8	71.1
7	Crandem$_{\log}$ (state+trans)	73.1	70.5	71.7
8	Crandem$_{\text{unnorm}}$ (state)	73.1	70.1	71.2
9	Crandem$_{\text{unnorm}}$ (state+trans)	73.1	70.6	71.8

a. System results using 61 phone class posteriors as input

	System	Dev	Core	Ext
1	Tandem (105 ftrs)	72.2	69.7	70.9
2	Tandem (48 ftrs)	72.5	70.2	71.2
3	CRF (state only)	72.7	70.3	71.4
4	CRF (state+trans)	72.7	70.9	71.6
5	MLP-Tandem	71.4	69.4	70.8
6	Crandem$_{\log}$ (state)	73.0	70.7	71.7
7	Crandem$_{\log}$ (state+trans)	73.4	71.2	72.4
8	Crandem$_{\text{unnorm}}$ (state)	72.9	70.6	71.7
9	Crandem$_{\text{unnorm}}$ (state+trans)	73.4	70.8	72.4

b. System results combining 61 phone class posteriors with 44 phonological feature posteriors

Table 1. Percent phone accuracies on TIMIT for development, core test, and extended test sets for the 9 systems outlined in Figure 1. Significance at the p$\leq$0.05 level is approximately 0.9%, 1.4%, and 0.6% percentage difference for these datasets, respectively.

We started by training two sets of MLP posterior estimates to be used by all of the systems. The first MLP predicts, at a frame-level, the posterior over 61 TIMIT phone classes from 13-dimensional PLP coefficients (with velocity and acceleration coefficients). We also trained a set of MLPs on the same data to predict phonological attribute values over 8 separate feature classes (sonority, consonant manner, place, and voicing, vowel height, frontness, rounding and laxness); this provided a set of 44 attribute posteriors to be used in parallel with the 61 phone classes.[2] Each MLP utilized 2000 hidden units and was trained over the TIMIT si and sx training sentences. Results from systems that use only the 61 phone posteriors are described in Table 1a, whereas the combined 61 and 44 posterior results are given in Table 1b.

As a baseline, we linearized the posteriors and performed a KL transformation in order to train a traditional Tandem system (System 1). We were also concerned that the CRF, which utilizes a 48-phone label set, may be performing a dimensionality reduction of the data; therefore we trained a Tandem system with the top 48 features from the KLT (System 2) for comparison. The HTK Toolkit [11] was used to train 32-mixture tied-triphone systems; testing for all HTK systems used an unweighted triphone lattice that enforces triphone constraints, which in our experience works better than bigram phone models on this task.

The CRF system based on our previous work [3] is represented by System 3; the posteriors are taken directly from the MLPs without linearization or whitening.[3] Decoding for this system and System 4 is performed using a Viterbi algorithm built into our CRF software. Unlike the HMM, the CRF models are one-state-per-phone monophone systems.[4] System 4 introduces posterior estimates as transition features as well as state features in the CRF (Section 2). Posterior estimates derived from these trained CRF systems give rise to Crandem$_{\log}$ (Systems 6,7) and Crandem$_{\text{unnorm}}$ systems (Systems 8,9).

One other question is how much we gain through the optimization of the joint sequence likelihood by the CRF, rather than using a local estimator. To address this issue, an MLP was trained on the same data as CRF system 3; posterior estimates from this MLP were used to train a Tandem system (labeled MLP-Tandem, System 5).

All HMM-based systems were tuned on a development set consisting of 400 utterances from the test set, as defined by Halberstadt and Glass [12]. Results are reported for the traditional "core" set of 192 utterances, as well as the 944 utterances distinct from the 400 development utterances (labeled the "extended", or "ext" set).

4. RESULTS AND DISCUSSION

Results from all systems are shown in Table 1. The differences in the various systems are not significant on the core test set (due to the small size of the core), so following common practice results are also reported for the extended test set and all measure of significance are reported as measured against this extended set using a one-tailed Z test between pairs of systems. Table 1a shows the aggregated results of all systems using only the 61 phone class posteriors as inputs, while Table 1b shows the aggregated results of all systems using both the 61 phone class posteriors and the 44 phonological feature posteriors as inputs.

As shown in Table 1a, when only the phone class posteriors are used, System 3 (CRF state features only) performs significantly worse than Systems 1&2 (Tandem systems).[5] However, this performance discrepancy disappears in System 4 (CRF state and transition features). In addition, when both phone class and phonological feature posteriors are used, as shown in Table 1b, there is no significant different between the Tandem systems and the CRF systems.

In almost all cases, the Crandem system performs significantly better than either its corresponding Tandem system or its corresponding CRF system. Specifically, Crandem Systems 7&9 significantly outperform Tandem Systems 1&2 and CRF Systems 3&4 regardless of the input features used. Crandem System 6 shows significant performance gains over its matching CRF System 3 only when the input is restricted to 61 phone class features. In all other cases, the small improvement of CRF System 6 over its corresponding CRF and Tandem systems is insignificant.

[2]For more information about the particular phonological feature definitions, please see [3].

[3]There are some minor differences with the system reported in [3]; chief among these is a new stochastic gradient training algorithm similar to that used in [2].

[4]We experimented with three-state-per-phone models in the CRF framework, but found that when posteriors were used as input, there was not much difference in performance. It is interesting to speculate that this is because of the context window in the MLP helping with durational constraints, but this is clearly unproven.

[5]This is contrary to our previous results where comparisons were between CRFs and an 8-mixture Gaussian HMM system; Systems 1 and 2 in this study feature 32 mixtures. Many more parameters are needed in the HMM to surpass the CRF performance.

System	Dev	Core	Ext
PLP + System 7b	74.3	71.8	73.3

Table 2. Percent phone accuracy for TIMIT with an HMM system trained with PLP coefficients appended to System 7b ($\text{Crandem}_{\log}$ (state+trans) trained on 61 phone class and 44 phonological attribute posteriors).

The gains in performance by the Crandem systems over their corresponding Tandem systems cannot be accounted for merely by dimensionality reduction, as can be seen by comparing the results of System 2 in each table to its corresponding Crandem systems. In both cases, the Crandem System 7 significantly outperforms even the dimensionality reduced Tandem System 2. In addition, as can be seen by comparing System 1&2 in each table, the performance improvement provided by dimensionality reduction is not significant for either Tandem system using these input features, suggesting that the dimensionality reduction performed by the CRF in the Crandem system is not the main cause for the improvement.

Comparing the results of the MLP combined Tandem system (System 5) with the various Crandem systems, we see that the Crandem systems always significantly outperform the corresponding Tandem system using inputs combined via a MLP. In both input cases, the results provided by combining features via joint estimation using CRFs outperform the result of combining features via the MLP local estimators; it is possible that this is due to the non-locality of the information integration of the CRF, or the conditional maximum likelihood training criterion of the CRF, but clearly more investigation is needed into this effect.

It is notable that the introduction of transition features to the CRF gives a small, consistent, but insignificant boost to accuracy when compared to CRF systems using only state-based features. Additionally, the use of CRF models with transition features gives a consistent, slightly significant boost to Crandem systems using these input features. This indicates that these results may carry over to more complex CRF models using richer feature sets. There does not seem to be a difference whether log posteriors or unnormalized posteriors are used in the Crandem system.

When comparing the results across feature sets, the CRF-based models benefit more by adding phonological feature posteriors than the pure Tandem-based models do. In all cases but one (System 8) the CRF and Crandem models show a small but significant improvement moving from using only the phone posteriors to using phone posteriors and phonological feature posteriors. Meanwhile, adding phonological features to Systems 1&2 does not significantly improve the performance in either case.

Finally, in line the experiments that combine Tandem features with traditional acoustic features [8], we took the best-performing development system (System 7b, a $\text{Crandem}_{\log}$ system that incorporates transition features and uses the combined phone/phonological feature input set) and concatenated those CRF posteriors with PLP features. In Table 2, one can see that there is a small but significant boost in phone accuracy over System 7b, as well as a significant gain over the original PLP features from which all of the other feature representations are derived.

5. CONCLUSIONS

These results mark a first step towards taking CRF models beyond phone recognition and into the realm of word recognition. The Crandem system methodology provides a means for taking results from a CRF estimator and integrating them into an existing HMM-based system. This will likely enable us to make better use of the growing number of CRF feature sets appearing in the field in order to improve our models of speech recognition. Our results show that this method can outperform both CRFs and Tandem systems for the task of phone recognition. Our work for the near future involves applying the Crandem model to the task of word recognition on corpora other than the TIMIT corpus, as well as investigating whether these results carry over to other CRF function definitions.

6. ACKNOWLEDGMENTS

The authors gratefully acknowledge support for this work provided in part by NSF CAREER grant IIS-0643901 and a Dayton Area Graduate Studies Institute / AFRL Student-Faculty Fellowship. The opinions expressed in this work are those of the authors and not of any funding agency.

7. REFERENCES

[1] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 18th International Conference on Machine Learning*, 2001.

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

[3] J. Morris and E. Fosler-Lussier, "Further experiments with detector-based conditional random fields in phonetic recognition," in *International Conference on Acoustic, Speech, and Signal Processing (ICASSP-2007)*, Honolulu, Hawaii, 2007.

[4] Y. H. Abdel-Haleem, *Conditional Random Fields for Continuous Speech Recognition*, Ph.D. thesis, University of Sheffield, Sheffield, UK, November 2006.

[5] C.-Y. Lin and H.-C. Wang, "Attribute-based Mandarin speech recognition using conditional random fields," in *Proc. Interspeech*, Antwerp, Belgium, 2007.

[6] Y. Wang, "Integrating phone boundary and phonetic boundary information into ASR sytems," M.S. thesis, The Ohio State University, Columbus, OH, 2007.

[7] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature stream extraction for conventional HMM systems," in *ICASSP*, 2000.

[8] Q. Zhu, A. Stolcke, B. Chen, and N. Morgan, "Using MLP features in SRI's conversational speech recognition system," in *Proc. Interspeech*, 2005, pp. 2141–2144.

[9] H.-K. J. Kuo and Y. Gao, "Maximum entropy direct models for speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, May 2006.

[10] F. Sha and F. Pereira, "Shallow parsing with Conditional Random Fields," in *Proc. of HLT, NAACL*, 2003.

[11] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge Unveristy Engineering Department, 2002, http://htk.eng.cam.ac.uk.

[12] A. K. Halberstadt and J. R. Glass, "Heterogeneous acoustic measurements for phonetic classification," in *Proceedings of Eurospeech*, 1997.