Society for Science and Education
United Kingdom

TMLAI Transactions on Machine Learning and Artificial Intelligence

Volume 3, Issue 2
ISSN 2054-7390

Frame Based Postprocessor for Speech Recognition Based on Augmented Conditional Random Fields

[1]Yasser Hifny

[1]*Faculty of computers and information systems, University of Helwan, Egypt;*

yhifny@fci.helwan.edu.eg

Abstract

In this paper, we present a novel postprocessor for speech recognition using the Augmented Conditional Random Field (ACRF) framework. In this framework, a primary acoustic model is used to generate state posterior scores per frame. These output scores are fed to the ACRF postprocessor for further frame based acoustic modeling. Since ACRF explicitly integrates acoustic context modeling, the postprocessor has the ability to discover new context information and to improve the recognition accuracy. The results on the TIMIT phone recognition task show that the proposed postprocessor can lead to significant improvements especially when Hidden Markov Models (HMMs) were used as primary acoustic model.

Keywords: Hidden Markov models; augmented conditional random fields; deep conditional random fields; speech recognition postprocessor.

1 Introduction

Acoustic modeling postprocessing based on methods derived from Conditional Random Fields [1] is an active area of research [2], [3], [4]. CRFs have a generic way to define feature functions (constraints). Consequently, the feature functions play a vital role in defining the model and its applications [5]. In this work, we present a frame based postprocessor for speech recognition based on ACRFs [6, 7]. The ACRFs paradigm is a nonlinear variant of CRFs where the feature functions are computed from scoring a large number of Gaussians. The projection of low dimensional acoustic data into a high dimensional (augmented) space aims to simplify the classification problem. The main advantage of this framework is that acoustic context information is explicitly integrated to handle the sequential phenomena of the speech signal and hence can be expected to improve the recognition accuracy. The ACRFs can be efficiently estimated using the *Approximate Iterative Scaling* (AIS) algorithm.

In the original ACRF framework, the process of augmenting the low dimensional space to obtain a high dimensional space ($\mathbf{o}_t \rightarrow \mathbf{o}_t^{\text{Aug}}$) is based on the following algorithm:

1. A large number of Gaussians is estimated from the training data using the EM algorithm [8].

2. The Gaussians provide scores for each frame.

DOI: 10.14738/tmlai.32.943
Publication Date: 14th April, 2015
URL: http://dx.doi.org/10.14738/tmlai.32.943

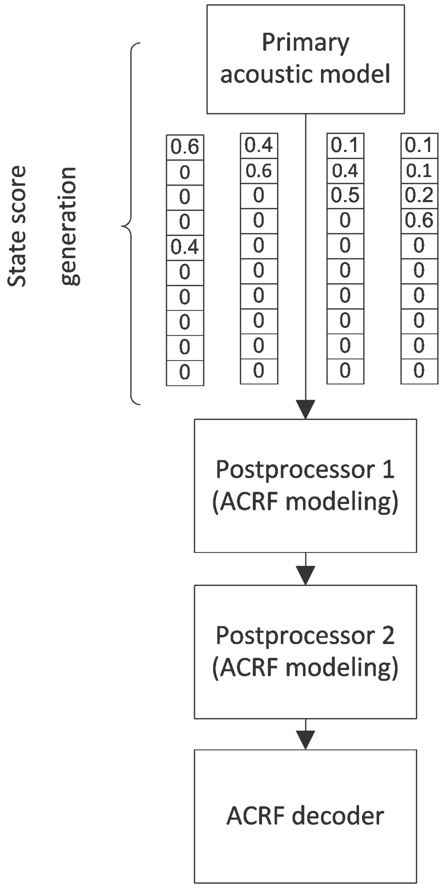


Figure 1: Frame based postprocessor using augmented conditional random field (ACRF) framework.

3. The scores are sorted and only the n-best scores are retained to reduce the the storage requirements during the training. Typically, the n-best nearest-neighbor shortlist size is set to 10.

4. An augmented vector is constructed and its size d^{Aug} equals the number of Gaussians in the recognition problem. A state feature value is calculated as a pruned posterior score for each Gaussian and is given by

$$b_i(\mathbf{o}_t) = \frac{\mathcal{N}_i(\mathbf{o}_t;\lambda)}{\sum_j \mathcal{N}_j(\mathbf{o}_t;\lambda)} \approx \frac{\mathcal{N}_i(\mathbf{o}_t;\lambda)}{\sum_{j\in n-\mathrm{best}} \mathcal{N}_j(\mathbf{o}_t;\lambda)}, \tag{1}$$

where $\mathcal{N}_i(\mathbf{o}_t;\lambda) \approx 0$ for $i \notin n-$best list and the normalization step is conceptually redundant to improve the ACRFs training speed.

Frame based acoustic models generate state scores. These state scores are fed to a decoder to generate the recognition hypothesis. For example, in HMMs [9, 10, 11, 12], an acoustic feature vector $\mathbf{o}_t$ may be generated, with an output probability density function $b_j(\mathbf{o}_t)$, which is associated with state j. A mixture of Gaussian distributions is typically used to model the output distribution for each state,

$$b_j(\mathbf{o}_t) = \sum_{m=1}^{M} c_{jm}\mathcal{N}(\mathbf{o}_t; \mu_{jm}, \Sigma_{jm}), \tag{2}$$

where M is the number of mixture components, c_{jm} is the component weight and $\sum_m^M c_{jm} = 1$. μ_{jm} and Σ_{jm} are the component specific mean vector and covariance matrix respectively. These state scores can be sorted and normalized in a similar way as in Equation (1). Hence, the normalized state scores is given by:

$$x_j(\mathbf{o}_t) = \frac{b_j(\mathbf{o}_t)}{\sum_{\mathbf{s}} b_{\mathbf{s}}(\mathbf{o}_t)} \approx \frac{b_j(\mathbf{o}_t)}{\sum_{k\in n-\text{best}} b_k(\mathbf{o}_t)}, \tag{3}$$

where $b_j(\mathbf{o}_t) \approx 0$ for $j \notin n-$best list.

The generated normalized state scores in Equation (3) are fed to ACRF postprocessor for further acoustic modeling. The ACRF output state scores can be normalized in a similar way and fed to a second layer ACRF for further acoustic modeling. An example of the described process is shown in Figure 1. By explicitly integrating acoustic context modeling using the ACRFs, the postprocessors have the ability to discover new context information and to improve the recognition accuracy. This is the main motivation behind the work.

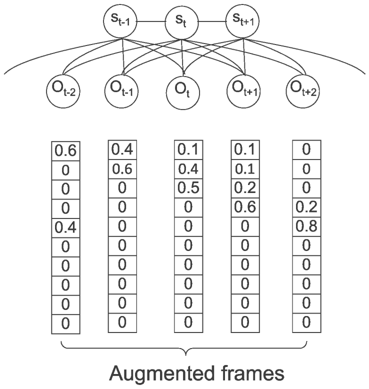
In this work, we investigated three different primary acoustic models which have different modeling power.[1] In particular, HMMs were tested as the main acoustic model. In addition, ACRF acoustic modeling as described in [7] was evaluated as a primary acoustic model. Finally, powerful deep conditional random fileds (DCRFs) [13] were developed as a primary acoustic model. DCRFs are a variant of hybrid deep neural networks DNN/HMM [14], [15],[16],[17],[18] formulated using the maximum entropy principle [19]. The main goal of testing different primary acoustic models is to show the modeling effect of using an ACRF postprocessor.

This paper is organized as follows: the mathematical formulation of ACRFs is given in Section 2. Section 3 describes how to compute the normalized state scores for different primary acoustic models. Experimental results on a phone recognition task are given in Section 4. Finally, a summary of the presented work is given in the conclusions.

2 Augmented Conditional Random Fields

ACRFs are undirected graphical models that maintain the Markov properties of HMMs. They operate in a high dimensional (augmented) space to improve the discrimination between speech classes. This augmented space is constructed by

[1] A primary acoustic model provides the features to the postprocessors.



0.6	0.4	0.1	0.1	0
0	0.6	0.4	0.1	0
0	0	0.5	0.2	0
0	0	0	0.6	0.2
0.4	0	0	0	0.8
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

Figure 2: ACRF phone model with state scores computed from a window of augmented frames.

scoring a large number of Gaussians. In addition, by using a large window of augmented frames, acoustic context information is explicitly integrated allowing the model to handle the sequential nature of speech signals. Hence, the HMM conditional independent assumptions are relaxed in this framework. ACRFs feature functions are based on pruned posterior scores to improve the training speed. The ACRFs have a batch training algorithm that scales for a large amount of training data.

The linear chain undirected graphical model behind the ACRF is shown in Figure 2. The model has the following properties:

- It obeys the Markovian property.
- The state scores are computed from the augmented frames (pruned posterior scores).

Given a state sequence $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_T)$ and a time sequence of speech frames or acoustic observations associated an utterance $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T)$, the maximum entropy conditional distribution defining ACRFs is

$$P_\Lambda(\mathbf{S}|\mathbf{O}) = \frac{1}{Z_\Lambda(\mathbf{O})} \prod_{t=1}^{T} \exp\Big(\lambda_{\mathbf{s}_t\mathbf{s}_{t-1}} a(\mathbf{s}_t, \mathbf{s}_{t-1}) + \sum_{u=t-c}^{t+c} \sum_{i=1}^{d^{\mathrm{Aug}}} \lambda_{\mathbf{s}_t}^{ui} x_i(\mathbf{o}_u)\Big), \quad (4)$$

where $\lambda_{\mathbf{s}_t}^{ui}$ and $\lambda_{\mathbf{s}_t\mathbf{s}_{t-1}}$ are associated with the feature functions $x_i(\mathbf{o}_u)$ and the transition functions $a(\mathbf{s}_t, \mathbf{s}_{t-1})$.[2] The feature functions $x_i(\mathbf{o}_t)$ are computed as in Equation (3) when HMMs are used as a primary acoustic model. The number of frames in the acoustic context window is $w = 2c+1$. $Z_\Lambda(\mathbf{O})$ (Zustandsumme) is a normalization coefficient referred to as the partition functions and is given

[2] $a(\mathbf{s}_t, \mathbf{s}_{t-1})$ is binary valued and can be used to specify the transition topology.

URL:http://dx.doi.org/10.14738/tmlai.32.943

by

$$Z_\Lambda(\mathbf{O}) = \sum_{\mathbf{S}} \prod_{t=1}^{T} \exp\Big(\lambda_{\mathbf{s}_t\mathbf{s}_{t-1}} a(\mathbf{s}_t, \mathbf{s}_{t-1}) + \sum_{u=t-c}^{t+c} \sum_{i=1}^{d^{\text{Aug}}} \lambda_{\mathbf{s}_t}^{ui} x_i(\mathbf{o}_u)\Big), \tag{5}$$

and it can be computed efficiently using the forward algorithm [1].

The feature functions $x_i(\mathbf{o}_t)$ are computed in a different way for other primary acoustic models. Section 3 will explain how to compute these feature functions for ACRFs and DCRFs acoustic models. In particular, Equation (11) and Equation (16) are used for primary acoustic models based on ACRFs and DCRFs respectively.

The primary acoustic decoding results are based on state scores. Compared to the primary system, the ACRF postprocessing sees next to the current set of state scores also those of the neighboring frames, allowing the integration of context information in the augmented space. It is worth to mention that when acoustic context information is not modeled (i.e. $c = 0$), the ACRF postprocessor and the primary acoustic model should lead to the same recognition results.

2.1 ACRF Optimization

For R training observations $\{\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_r, \ldots, \mathbf{O}_R\}$ with corresponding transcriptions $\{W_r\}$, ACRFs are trained using the conditional maximum likelihood (CML) criterion to maximize the posterior probability of the correct word sequence given the acoustic observations. Exact lower bound optimization algorithms for CRFs are very slow [1]. Therefore, we use the *Approximate Iterative Scaling* (AIS) algorithm to speed up the training process. The value of the learning rate is the main difference between exact and approximate algorithms. An AIS algorithm update equation is given by:

$$\lambda_{ji}^{\tau+1}(\mathbf{O}) = \lambda_{ji}^{\tau}(\mathbf{O}) + \eta_{\text{AIS}} \log \frac{\mathcal{C}_{ji}^{\text{num}}(\mathbf{O})}{\mathcal{C}_{ji}^{\text{den}}(\mathbf{O})}, \tag{6}$$

where $\eta_{\text{AIS}} = \frac{1}{w}$ is called the *learning rate* and τ is the iteration number. The sparse accumulators of the sufficient statistics, $\mathcal{C}_{ji}(\mathbf{O})$, for the j^{th} state and i^{th} constraint are calculated as follows:

$$\mathcal{C}_{ji}^{\text{num}}(\mathbf{O}) = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_j^r(t|\mathcal{M}^{\text{num}}) \mathbf{o}_{rti}^{\text{Aug}}, \tag{7}$$

$$\mathcal{C}_{ji}^{\text{den}}(\mathbf{O}) = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_j^r(t|\mathcal{M}^{\text{den}}) \mathbf{o}_{rti}^{\text{Aug}}, \tag{8}$$

where r is the utterance index and $\mathbf{o}_t^{\text{Aug}} = [\mathbf{o}_{t-c}, \ldots, \mathbf{o}_t, \ldots, \mathbf{o}_{t+c}]$. Given the forward score $\alpha_j(t)$ and backward score $\beta_j(t)$, the occupation probability of being in state j at time t, γ_j, is given by:

$$\gamma_j(t|\mathcal{M}) = P(\mathbf{s}_t = j|\mathbf{O}; \mathcal{M}) = \frac{\alpha_j(t|\mathcal{M})\beta_j(t|\mathcal{M})}{Z_\Lambda(\mathbf{O}|\mathcal{M})}, \tag{9}$$

and to avoid the necessity of building lattices, the $\gamma_j(t|\mathcal{M})$ is approximated with state estimates as follows [20]:

$$\gamma_j(t|\mathcal{M}) = \frac{\exp\left(\sum_{u=t-c}^{t+c}\sum_{i=1}^{d^{\text{Aug}}} \lambda_j^{ui} x_i(\mathbf{o}_u)\right)}{\sum_{\mathbf{s}} \exp\left(\sum_{u=t-c}^{t+c}\sum_{i=1}^{d^{\text{Aug}}} \lambda_{\mathbf{s}}^{ui} x_i(\mathbf{o}_u)\right)}. \tag{10}$$

3 State scores generation

Three different primary acoustic models which have different modeling power were developed in this work. For HMM, the generated normalized state scores are computed as in Equation (3) . For ACRF and DCRFs, the goal of this section is to show how to compute their normalized state scores.

3.1 ACRF as a primary acoustic model

ACRFs can be used as a primary acoustic model if the input features to ACRFs are based on Equation (1) . The parameter estimation is exactly identical to described in Section 2. The normalized state scores are given by

$$x_j(\mathbf{o}_t) = \frac{\exp\left(\sum_{u=t-c}^{t+c}\sum_{i=1}^{d^{\text{Aug}}} \lambda_j^{ui} b_i(\mathbf{o}_u)\right)}{\sum_{\mathbf{s}} \exp\left(\sum_{u=t-c}^{t+c}\sum_{i=1}^{d^{\text{Aug}}} \lambda_{\mathbf{s}}^{ui} b_i(\mathbf{o}_u)\right)} \approx \frac{\exp\left(\sum_{u=t-c}^{t+c}\sum_{i=1}^{d^{\text{Aug}}} \lambda_j^{ui} b_i(\mathbf{o}_u)\right)}{\sum_{k\in n-\text{best}} \exp\left(\sum_{u=t-c}^{t+c}\sum_{i=1}^{d^{\text{Aug}}} \lambda_k^{ui} b_i(\mathbf{o}_u)\right)}, \tag{11}$$

where $\exp\left(\sum_{u=t-c}^{t+c}\sum_{i=1}^{d^{\text{Aug}}} \lambda_j^{ui} b_i(\mathbf{o}_u)\right) \approx 0$ for $j \notin n-$best list.

3.2 DCRF as a primary acoustic model

Training CRFs on the top of a hidden layer constructed from scoring a large number of sigmoid functions was introduced in [17]. One way to improve this approach is to compute the state scores based on a DNN that has several hidden layers [21]. Deep Conditional Random Fields acoustic models are a particular implementation of linear chain CRFs where the state scores are computed based on a DNN that has several hidden layers [13]. The output layer of DCRFs is based on linear activation functions while in hybrid DNN/HMM it is based on softmax activation functions. This is the main difference between DCRFs and hybrid DNN/HMM systems. A graphical representation of the DCRF acoustic model is shown in Figure 3. The conditional distribution defining DCRFs is given by

$$P_\Lambda(\mathbf{S}|\mathbf{O}) = \frac{1}{Z_\Lambda(\mathbf{O})}\prod_{t=1}^{T} \exp\Big(\lambda_{\mathbf{s}_t\mathbf{s}_{t-1}} a(\mathbf{s}_t, \mathbf{s}_{t-1}) + b_{\mathbf{s}_t}(\mathbf{o}_t)\Big), \tag{12}$$

where $b_{\mathbf{s}_t}(\mathbf{o}_t)$ is computed from a DNN scorer.

The feed-forward phase of a DNN scorer updates the output value of each hidden unit. Each hidden unit output is computed as follows:

$$\mathbf{o}_{tj}^{h} = \mathrm{sigm}(\sum_{i=1}^{n} \lambda_{ij} \mathbf{o}_{ti}^{h-1}), \tag{13}$$

where $\mathbf{o}_t^h$ is an output of a hidden layer, n is the number of inputs, and h is an index to a hidden layer. The sigmoid function is computed as follows:

$$\mathrm{sigm}(x) = \frac{1}{1+e^{-x}}. \tag{14}$$

The output of an hidden layer is forwarded to the next layer until the output layer is computed as follows (linear activation):

$$\mathbf{o}_{tj}^{N} = \sum_{i=1}^{n} \lambda_{ij} \mathbf{o}_{ti}^{N-1}, \tag{15}$$

where N is the index of the output layer. Hence, $b_{\mathbf{s}_t}(\mathbf{o}_t) = \mathbf{o}_{ts_t}^{N}$ connects a DNN scorer to CRFs.

The normalized state scores are given by

$$x_j(\mathbf{o}_t) = \frac{\exp\left(\mathbf{o}_{tj}^{N}\right)}{\sum_{\mathbf{s}} \exp\left(\mathbf{o}_{t\mathbf{s}}^{N}\right)} \approx \frac{\exp\left(\mathbf{o}_{tj}^{N}\right)}{\sum_{k\in n-\mathrm{best}} \exp\left(\mathbf{o}_{tk}^{N}\right)}, \tag{16}$$

where $\exp\left(\mathbf{o}_{tj}^{N}\right) \approx 0$ for $j \notin n-$best list.

4 Experiments

In this section, the standard TIMIT phone recognition task is used to evaluate the proposed approach described in this paper. The training sets consist of 462 speakers and results are computed using the 24 speaker core test set. The DNN training development set is based on 50 speakers from the test set [22]. The SA1 and SA2 utterances were not used.

The speech was analyzed using a 25ms Hamming window with a 10 ms fixed frame rate. The speech is represented using 12 mel frequency cepstral coefficients (MFCCs), energy, along with their first and second temporal derivatives, resulting in a 39 element feature vector. Another representation used for DCRFs is based on using a Fourier-transform-based filter-bank with 40 coefficients (plus energy) distributed on a mel-scale, together with their first and second temporal derivatives resulting in a 123 element feature vector. The features are pre-processed to have zero mean and unit variance and acoustic context information is integrated using a window of 9 frames (4 left + current frame+ 4 right) to construct the final frames.

The original 61 phone classes in TIMIT were mapped to a set of 48 labels, which were used for training [23]. After decoding, this set of 48 phone classes

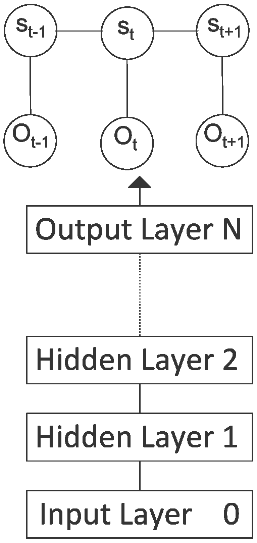


Figure 3: Linear chain DCRF model (the state scores are computed from a DNN).

was mapped down to a set of 39 classes. The phone error rate (PER) metric, which is analogous to word error rate, is used to report phone recognition results.

Each phone of the baseline HMMs was represented using a three state left-to-right model. Mixtures of Gaussian densities with diagonal covariance matrices were used for state scoring (emission probabilities). The HMMs were trained by the maximum likelihood criterion using the conventional EM algorithm [24]. Discriminative training based on Minimum Phone Error (MPE) criterion was used to refine the HMMs [25]. The acoustic scale was set to 1/6 and I-smoothing parameter τ was set to 100.

Similar to the HMMs, the ACRF-based models emply three-state left-to-right phone models. The transition parameters were initialized from trained HMM models. Other parameters were initialized to zero. The same model structure was used for postprocessor ACRFs. A Viterbi pass (forced alignment) of the reference transcription using HMMs trained using the maximum likelihood criterion was used to accumulate the $\mathcal{M}^{\text{num}}$ sufficient statistics. The number of frames in the acoustic context window, $w = 2c + 1$, was set to 19. For ACRFs primary acoustic models, 7917 Gaussians were used to construct the augmented space.

A powerful primary acoustic model based on DCRFs was evaluated. Each phone was represented using a three state left-to-right DCRF. The transition parameters were initialized from trained HMM models as in ACRFs. The DNN parameters were initialized to random values. The DNN has nine hidden layers and each layer has 2048 neurons. For training DCRFs, the PDNNTK toolkit [26] in combination with the Theano library [27] is used, allowing transparent

Table 1: HMM decoding results on TIMIT recognition task in terms of PER.

Model	10 Mix	40 Mix
HMM baseline	32.3%	29.9%
ACRF postprocessor1	28.7%	27.9%
ACRF postprocessor2	28.2%	27.5%

Table 2: Decoding results on TIMIT recognition task in terms of PER for different primary acoustic models.

Model	ACRFs	DCRFs
baseline	27.3%	22.7%
ACRF postprocessor1	26.7%	22.3%
ACRF postprocessor2	26.6%	22.5%

computation for CPUs and GPUs.

The acoustic modeling process starts with generating the state scores of the primary models in pruned posterior forms. These scores are fed to the first ACRF postprocessor. The output state scores of the first ACRF postprocessor are generated in pruned posterior forms and are fed to the second ACRF postprocessor in all experiments.

A generic bi-gram in-house decoder is used to generate the recognition phone sequence for the different acoustic models. Table 1 shows the decoding results when HMMs are used as a primary acoustic model. The results show that the first stage of ACRFs postprocessing leads to significant improvement in terms of PER. When ACRFs and DCRFs were used as primary acoustic models, the improvements are smaller than HMMs as shown in Table 2. The second stage of postprocessing did not lead to improvements. These results may suggest that ACRF postprocessing has limited ability for powerful acoustic models.

5 Conclusions

In this paper, an augmented conditional random field postprocessor for speech recognition is presented. In this framework, a primary acoustic model is used to generate state posterior scores per frame. These posterior scores are then used as input to an ACRF. The main goal of this process is to model the acoustic context information in a high dimensional space constructed using the primary acoustic model state scores. Consequently, the postprocessor acoustic model discovers new context information and improves the recognition accuracy. Three different primary acoustic models were investigated in this work (HMM, ACRF, and DCRF). Results on the TIMIT phone recognition task show that the proposed postprocessor can lead to significant improvements especially when HMMs were used as a primary acoustic model.

References

[1] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proc. ICML, 2001, pp. 282–289.

[2] M. Layton, M. Gales, Augmented statistical models for speech recognition, in: Proc. IEEE ICASSP, Vol. 1, France, 2006, pp. 129– 132.

[3] J. Morris, E. Fosler-Lussier, Conditional random fields for integrating local discriminative classifiers, Audio, Speech, and Language Processing, IEEE Transactions on 16 (3) (2008) 617–628. doi:10.1109/TASL.2008.916057.

[4] G. Zweig, P. Nguyen, D. V. Compernolle, K. Demuynck, L. Atlas, P. Clark, G. Sell, M. Wang, F. Sha, H. Hermansky, D. Karakos, A. Jansen, S. Thomas, G. Sivaram, S. Bowman, J. Kao, Speech recognition with seg- mental conditional random fields: A summary of the J HU CLSP summer workshop, in: Proc. IEEE ICASSP, 2011.

[5] M. Gales, S. Watanabe, E. Fosler-Lussier, Structured discriminative models for speech recognition, IEEE Signal Processing Magazine.

[6] Y. Hifny, Conditional random fields for continuous speech recognition, Ph.D. thesis, University Of Sheffield (2006).

[7] Y. Hifny, S. Renals, Speech recognition using augmented conditional random fields, IEEE Transactions on Audio, Speech and Language Processing 17 (2) (2009) 354-365

[8] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society 39 (1) (1977) 1–38.

[9] L. R. Rabiner, A tutorial on hidden Markov models and selected applica-tions in speech recognition, Proc. of IEEE 77 (2) (1989) 257–286.

[10] F. Jelinek, Statistical Methods for Speech Recognition, MIT Press, 1997.

[11] X. Huang, A. Acero, H.-W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm and System Development, Prentice Hall, 2001.

[12] J. Bilmes, What HMMs can do, IEICE Transactions on Information and Systems E89-D (3) (2006) 869–891.

[13] Y. Hifny, Acoustic modeling based on deep conditional random fields, Deep Learning for Audio, Speech and Language Processing, ICML.

[14] S. Renals, N. Morgan, H. Bourlard, M. Cohen, H. Franco, Connectionist probability estimators in HMM speech recognition, IEEE Transactions on Speech and Audio Processing.

[15] N. Morgan, H. Bourlard, Continuous speech recognition: An introduction to the hybrid HMM/connectionist approach, IEEE Signal Processing Mag- azine 12 (3) (1995) 25–42.

[16] B. Kingsbury, Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling, in: Proc. IEEE ICASSP, 2009, pp. 3761–3764. doi:10.1109/ICASSP.2009.4960445.

[17] R. Prabhavalkar, E. Fosler-Lussier, Backpropagation training for multilayer conditional random field based phone recognition, in: Proc. IEEE ICASSP, Vol. 1, France, 2010, pp. 5534 – 5537.

[18] G. Hinton, L. Deng, D. Yu, G. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, , B. Kingsbury, Deep Neural Networks for acoustic modeling in speech recognition, IEEE Signal Processing Magazine.

[19] E. T. Jaynes, On the rationale of maximum-entropy methods, Proc. of IEEE 70 (9) (1982) 939–952.

[20] Y. Hifny, S. Renals, N. Lawrence, A hybrid MaxEnt/HMM based ASR system, in: Proc. INTERSPEECH, Lisbon, Portugal, 2005, pp. 3017–3020.

[21] A. Mohamed, D. Yu, L. Deng, Investigation of full-sequence training of Deep Belief Networks for speech recognition, in: Interspeech, 2010.

[22] A. Halberstadt, J. Glass, Heterogeneous measurements and multiple classifiers for speech recognition, in: Proc. ICSLP, Vol. 3, Sydney, Australia, 1998, pp. 995–998.

[23] K.-F. Lee, H.-W. Hon, Speaker-independent phone recognition using hidden Markov models, IEEE Transactions on Speech and Audio Processing 37 (11) (1989) 1641–1648.

[24] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland, The HTK Book, Version 3.1, 2001.

[25] D. Povey, Discriminative training for large vocabulary speech recognition, Ph.D. thesis, University of Cambridge (2004).

[26] Y. Miao, PDNN: Yet Another Python Toolkit for Deep Neural Networks. URL `http://www.cs.cmu.edu/ ymiao/pdnntk.html`

[27] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, Y. Bengio, Theano: a CPU and GPU math expression compiler, in: Proceedings of the Python for Scientific Com- puting Conference (SciPy), 2010, oral Presentation.

Citing Information:

Yasser Hifny; Frame Based Postprocessor for Speech Recognition Based on Augmented Conditional Random Fields, Transactions on Machine Learning and Artificial Intelligence, Volume 3 No 2 April (2015); pp: 24-33

URL:http://dx.doi.org/10.14738/tmlai.32.943