Speech Recognition Using Augmented Conditional Random Fields

Yasser Hifny and Steve Renals, *Member, IEEE*

***Abstract*—Acoustic modeling based on hidden Markov models (HMMs) is employed by state-of-the-art stochastic speech recognition systems. Although HMMs are a natural choice to warp the time axis and model the temporal phenomena in the speech signal, their conditional independence properties limit their ability to model spectral phenomena well. In this paper, a new acoustic modeling paradigm based on augmented conditional random fields (ACRFs) is investigated and developed. This paradigm addresses some limitations of HMMs while maintaining many of the aspects which have made them successful. In particular, the acoustic modeling problem is reformulated in a data driven, sparse, augmented space to increase discrimination. Acoustic context modeling is explicitly integrated to handle the sequential phenomena of the speech signal. We present an efficient framework for estimating these models that ensures scalability and generality. In the TIMIT phone recognition task, a phone error rate of 23.0% was recorded on the full test set, a significant improvement over comparable HMM-based systems.**

***Index Terms*—Augmented conditional random fields (ACRFs), augmented spaces, discriminative compression, hidden Markov models (HMMs).**

I. Introduction

STATE-of-the-art automatic speech recognition systems use hidden Markov models (HMMs) [1]–[4] to model the temporal variation, with local spectral variability modeled using flexible distributions such as mixtures of Gaussian densities. HMMs can divide the acoustic space into a large number of small dense regions, assigning these regions to a large number of labels, or states, a process that is not unlike (soft) vector quantization and directly related to the definition of a pattern classification problem. Generative HMMs are well understood models and may be trained efficiently using the expectation-maximization (EM) algorithm [5]. Using Bayes rule, the coarse density estimates provided by HMMs can be used for discrimination. Consequently, HMMs provide a means to learn and generate spectral information in order to discriminate between speech classes.

HMMs trained using the EM algorithm maximize the likelihood of the data given the underlying parameterized distributions. If the true distribution that generated the data is indeed an HMM, then, given sufficient data, Bayes classification based on HMMs estimated using maximum likelihood will minimize the probability of classification error [6]. In practice, the decision boundaries constructed after generative training are not optimal and generative HMMs are not guaranteed to minimize the classification error. One way to address this problem within the HMM framework is to utilize the parameters efficiently to improve the discrimination between speech classes via discriminative training for HMMs [7]–[12].

Large-vocabulary continuous speech recognition systems based on continuous Gaussian mixture HMMs are very successful [13], mainly because the associated algorithms are computationally very efficient and scale well as the amount of training data increases. These attractive properties arise from two assumptions that lead to tractable inference and decoding. First, the Markov assumption enables the probability of the hidden state sequence $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_T)$ given a model $\mathcal{M}$ to be approximated using a first order Markov chain

$$P(\mathbf{S} \mid \mathcal{M}) = \prod_{t=1}^{T} P(\mathbf{s}_t \mid \mathbf{s}_{t-1}). \tag{1}$$

The second assumption is that of *conditional independence*, whereby the probability of an observation sequence, $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T)$, given a state sequence $\mathbf{S}$ and a model $\mathcal{M}$ is assumed to be

$$p(\mathbf{O} \mid \mathbf{S}, \mathcal{M}) = \prod_{t=1}^{T} p(\mathbf{o}_t \mid \mathbf{s}_t). \tag{2}$$

The conditional independence assumption is problematic since the slowly varying articulatory system produces long range interframe correlations. There is a mutual relationship between feature extraction from the speech signal and acoustic modeling based on HMMs. An ideal feature extraction method for speech recognition would find a set of compact features representing the observation space, while preserving the information needed to discriminate between speech classes. These limitations of the HMM may be addressed in part through the use of linear [14]–[16] or nonlinear feature projection methods [17]–[22], which extract new sufficient statistics that take into account acoustic context and improve the discrimination between speech classes. This may be achieved without changing the underlying HMM framework.

A discrete state space formulation, typically an HMM, is used for sequential modeling in speech recognition. There have been a number of significant enhancements to the underlying formulation that may be grouped into three areas, two related to pattern classification and one to sequential processing.

Manuscript received May 25, 2007; revised October 03, 2008. Current version published January 14, 2009. The work of H. Hifny was supported in part by a Motorola studentship at the University of Sheffield. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mari Ostendorf.

Y. Hifny is with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: yhifny@us.ibm.com).

S. Renals is with the Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9LW, U.K. (e-mail: s.renals@ed.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2008.2010286

1558-7916/$25.00 © 2009 IEEE

Authorized licensed use limited to: The University of Edinburgh. Downloaded on February 10, 2009 at 12:51 from IEEE Xplore. Restrictions apply.

- Augmenting the *state space* by increasing the number of hidden states. This can done by using context-dependent phone models which use a window of left and right neighboring phones [23]–[27].
- Augmenting the *observation space* with a large number of dimensions, which can simplify the classification problem [28], [29].
- Relaxing the HMM conditional independence assumptions, which can be done by integrating acoustic context information in the modeling process to take into account longer time intervals [30]–[32]. Acoustic context information may be incorporated using dynamic features [33] or implicitly based on feature projection [20], [21].

In this paper, a new acoustic model closely related to the HMM framework is proposed and evaluated. This framework focuses on augmenting the observation space and integrating the acoustic context information, thus relaxing the HMM conditional independence assumptions. Augmenting the state space is a well established idea in acoustic modeling research [23]–[27] and is not addressed in this work. Hence, the main motivation and our goal is to improve the discrimination between speech classes by formulating the acoustic modeling problem in a high-dimensional (augmented) space and explicitly integrating acoustic context information.

Augmented conditional random fields (ACRFs) are flexible acoustic models specifically designed to take advantage of context information in an augmented space. Unlike a low-dimensional HMM formulation (typically 40–100 dimensions), the ACRF formulation (typically 10^6 dimensions) will create acoustic models with large numbers of parameters estimated from data. Consequently, the ACRF formulation poses many research problems and raises issues about scaling to large amounts of training data.

In Section II, augmented spaces, which are sparse, high-dimensional acoustic spaces, are proposed and developed. The ACRF graphical model and its conditional distribution are detailed in Section III. Section IV describes the approximate iterative scaling (AIS) algorithm, which is used to train ACRFs. The AIS algorithm relies on particular properties of the ACRF to improve the speed of training. Moving to high-dimensional spaces may limit the scalability of this approach, so a discriminant compression algorithm is proposed (Section V), which allows the system to integrate the acoustic context in the augmented spaces and the parameter space to be pruned without additional computational cost during the training process. Experiments on the TIMIT phone recognition task are presented in Section VI. Finally, Section VII discusses further the behavior of ACRFs, and establishes connections to related models.

II. Augmented Spaces

Feature projection into high-dimensional spaces is a powerful tool to simplify classification problems, since high-dimensional spaces are more likely to be linearly separable than low-dimensional spaces [34], as illustrated in Fig. 1. This is usually achieved by mapping the low-dimensional input space into a high-dimensional space, with linear decision boundaries used for classification.[1] The dimensionality of a kernel space is N, where N is the total number of frames and its construction time or storage complexity is $O(N^2)$. In some cases, it is computationally more efficient to focus on dense regions in the observation space and construct spaces with dimensionality $\ll N$. We refer to such spaces as *Augmented Spaces*.

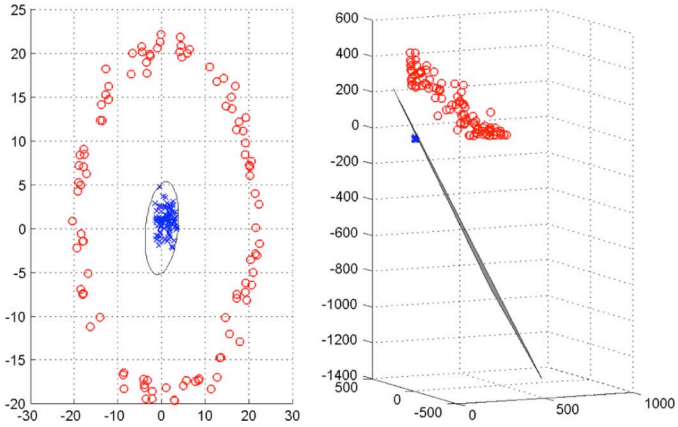

Fig. 1. Two-dimensional classification problem with nonlinear decision boundary is linearly separable in three-dimensional space with a transformation function ϕ: $\mathbb{R}^2 \rightarrow \mathbb{R}^3 = (\mathbf{o}_1, \mathbf{o}_2) \rightarrow (\mathbf{o}_1^2, \mathbf{o}_1\mathbf{o}_2, \mathbf{o}_2^2)$.

Two augmentation steps are implemented in this work to construct an augmented space.

- Augmentation based on the construction of high-dimensional spaces using a large number of constraint functions for each acoustic vector (i.e., $\mathbf{o}_t \rightarrow \mathbf{o}_t^{\text{Aug}}$).
- To take advantage of acoustic context, we add the surrounding frames to the current frame to have further augmentation (i.e., the discrimination between states will be a function, $\hat{f}(\mathbf{o}_{t-c}^{\text{Aug}}, \ldots, \mathbf{o}_{t+c}^{\text{Aug}})$, in the acoustic context).

The dimensionality of the resultant augmented space is very high (typically 10^6 dimensions) but if the constraint functions are chosen such that most elements of observed vectors in the augmented space are close to zero, then they may be pruned and the effective dimensionality of the problem will be substantially reduced. The two steps of the augmentation process are detailed below.

A. Augmentation by Parametric Constraints

The definition of the constraint functions is dependent on prior knowledge of the particular application. These constraints, with parameters λ, are used to locate dense regions in observation spaces with arbitrary resolution. The process of augmenting the low-dimensional space to result in a high-dimensional space $\mathbf{o}_t \rightarrow \mathbf{o}_t^{\text{Aug}}$, starts with the application of a large number of constraints $g_i(\mathbf{o}_t; \lambda)$ to the observed data. Then, the constraints are sorted according to their scores and only the n-best are retained.[2] The indices of the top n constraints represent the kernel functions that are most responsible for activating the acoustic frame under consideration (i.e., the acoustic regions most likely to account for the current frame). The selection of an n-best

[1]More recently, approaches that employ feature spaces induced by Mercer's kernels [35], [29] have been widely used since they are theoretically attractive, enabling computations in possibly infinite-dimensional feature spaces to be performed in finite-dimensional kernel spaces.

[2]Typically, the n-best nearest-neighbor shortlist size is set to 10.

Authorized licensed use limited to: The University of Edinburgh. Downloaded on February 10, 2009 at 12:51 from IEEE Xplore. Restrictions apply.

shortlist is essential to reduce the storage requirements of the approach. Once such a shortlist is available, the augmented vector is constructed and its size d^{Aug} equals the number of constraints in the recognition problem. A state constraint value in the new augmented space is calculated as a pruned posterior score for each parametric constraint and is given by

$$b_i(\mathbf{o}_t, \mathbf{s}_t) = \frac{g_i(\mathbf{o}_t; \lambda)}{\sum_j g_j(\mathbf{o}_t; \lambda)} \approx \frac{g_i(\mathbf{o}_t; \lambda)}{\sum_{j \in n-\text{best}} g_j(\mathbf{o}_t; \lambda)} \quad (3)$$

where the normalization step is conceptually redundant (discussed later). In general, there are several choice for the form of the parametric constraints but we are interested in any exponential family (*e-family*) activation functions or densities since their scores are positive. In this paper, we use diagonal Gaussian density functions, estimated using the EM algorithm to locate the dense regions

$$g_i(\mathbf{o}_t; \lambda) = p_i(\mathbf{o}_t \mid \theta) = \mathcal{N}(\mathbf{o}_t; \mu_i, \Sigma_i). \quad (4)$$

The resulting Gaussians are used to estimate the likelihood score for an observation, and a normalized version of that likelihood score will take the role of the constraint posterior score (3) in the augmented spaces framework. Scoring a large number of Gaussians may be accelerated using Gaussian selection techniques [36], [37].

The augmented spaces framework supports other e-family activation functions. Samples of these activation functions are

$$g_i(\mathbf{o}_t; \lambda) = \exp\left(-\frac{\|\mathbf{o}_t - \mu_i\|^2}{2\sigma_i^2}\right) \quad (5)$$

$$g_i(\mathbf{o}_t; \lambda) = \exp\left(\mathbf{o}_t^T \Lambda_i \mathbf{o}_t + \lambda_i^T \mathbf{o}_t + b_{i0}\right) \quad (6)$$

$$g_i(\mathbf{o}_t; \lambda) = \exp\left(\lambda_i^T \mathbf{o}_t + b_{i0}\right). \quad (7)$$

The e-family activation functions (5) and (6) can be estimated by accumulating up to second order statistics resulting in *quadratic* discriminant functions. The e-family activation function (7) is based on *linear* discriminant functions estimated from first-order statistics.

The augmentation process $\mathbf{o}_t \rightarrow \mathbf{o}_t^{\text{Aug}}$ is sketched in Fig. 2. The dense regions, those regions where most data points are projected, are defined using the hyperellipsoids derived from the eigen decomposition of the covariance matrix in case of the Gaussian activation. In addition, the orientation of the hyperellipsoid axes associated with diagonal covariance Gaussians are parallel to the coordinate axes. The pruned posterior scores are obtained for Gaussians near the point X. Consequently, most of the elements of an augmented vector $\mathbf{o}_t^{\text{Aug}}$ are zero as they are considered outliers for the point X. Moreover, the sum of the elements in an augmented vector is equal to 1.0. Finally, the dimensionality of the augmented space is equal to the number of Gaussians.

Although the augmented vector dimensionality is very high, few elements of the augmented vector are nonzero. These elements represent the effective dimensionality of the augmented vector. For example, if we construct an augmented space with an augmented dimensionality $d^{\text{Aug}} = 2\,000\,000$ and an n-best shortlist size of 10, the effective dimensionality $d^{\text{Eff}} = 10$.

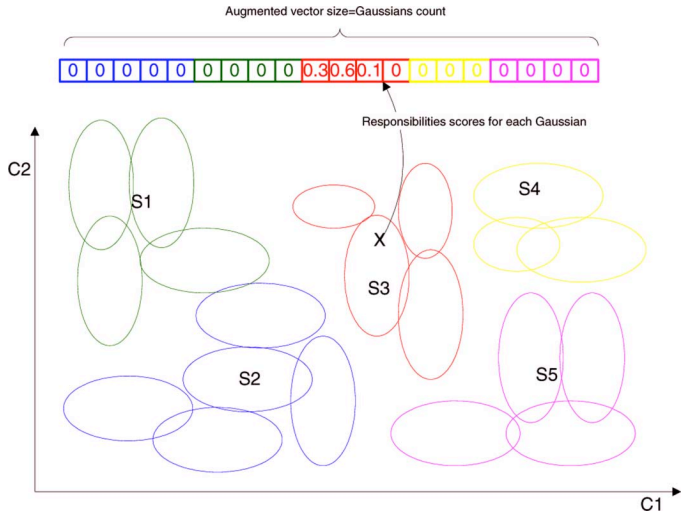

Fig. 2. Two-dimensional space is partitioned into 20 regions specified by diagonal Gaussians. The augmented space dimensionality is equal to the number of Gaussians (constraint functions) in the acoustic space. Hence, the two dimension space is augmented to a 20-dimensional space $\mathbb{R}^2 \rightarrow \mathbb{R}^{20}$. The augmented vector is constructed by calculating a posterior score for each Gaussian. The majority of the elements of the augmented vector have very low posterior score (i.e., $\approx$ zero). In this case, the n-best shortlist size is set to 3.

B. Augmentation by Adding the Acoustic Context

Acoustic context, taking into account a longer time interval for state discrimination within the modeling process, has been used previously in hybrid connectionist/HMM acoustic models [38] and for discriminant feature extraction [39], [18], [19]. Once the high-dimensional augmented vector $\mathbf{o}^{\text{Aug}}$ is constructed, it is possible to take advantage of acoustic context by adding the surrounding augmented frames $\mathbf{o}_{t-c}^{\text{Aug}}, \ldots, \mathbf{o}_{t+c}^{\text{Aug}}$ to the current frame during state scoring. As discussed below, this approach is well matched to the ACRF framework.

Using context modeling leads to a minor change to the computational complexity of augmented space construction, but adding these surrounding frames to a sparse augmented vector increases the problem dimensionality to wd^{Aug} and its effective dimensionality to wd^{Eff}, where $w = (2c+1)$

III. Augmented Conditional Random Fields (ACRFs)

ACRFs incorporate acoustic context information into an augmented space in order to model the sequential phenomena of the speech signal. ACRFs are derived from linear chain CRFs[3] [40], which are undirected graphical models that maintain the Markov properties of HMMs, formulated using the maximum entropy (MaxEnt) principle [41]. Linear chain CRFs can be thought as the undirected graphical twins for HMMs regardless of their training (generative or discriminative). ACRF acoustic models are a particular implementation of linear chain CRFs developed for augmented acoustic spaces. A graphical representation of the ACRF acoustic model is shown in Fig. 3; its states are dependent on arbitrary acoustic observations. The conditional independence properties of the HMM are relaxed explicitly in

[3]More precisely, ACRFs are a nonlinear form of linear chain CRFs.

Authorized licensed use limited to: The University of Edinburgh. Downloaded on February 10, 2009 at 12:51 from IEEE Xplore. Restrictions apply.

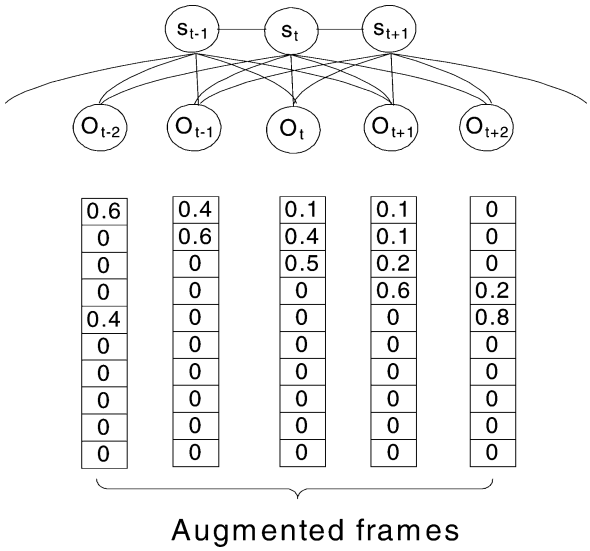

Fig. 3. ACRF model for phone representation that is dependent on arbitrary acoustic observations. Modeling the acoustic context by concatenating a window of multiple acoustic frames is equivalent to having a state dependence on the previous/following c frames ($c = 2$ in this ACRF model).

the ACRF acoustic model with the sufficient statistics collected from an augmented space.

The conditional distribution defining ACRFs is given by

$$P_\Lambda(\mathbf{S}\,|\,\mathbf{O}) = \frac{1}{Z_\Lambda(\mathbf{O})}\prod_{t=1}^{T}\exp\left(\lambda_{\mathbf{s}_t\mathbf{s}_{t-1}}a(\mathbf{s}_t,\mathbf{s}_{t-1}) + \sum_{u=t-c}^{t+c}\sum_{i=1}^{d^{\mathrm{Aug}}}\lambda_{\mathbf{s}_t}^{ui}b_i(\mathbf{o}_u,\mathbf{s}_t)\right) \quad (8)$$

where

- $P_\Lambda(\mathbf{S}\,|\,\mathbf{O})$ obeys the Markovian property $P_\Lambda(\mathbf{s}_t\,|\,\{\mathbf{s}_j\}_{j\neq t},\mathbf{O}) = P_\Lambda(\mathbf{s}_t\,|\,\mathbf{s}_{t-1},\mathbf{O})$.
- $\lambda_{\mathbf{s}_t}^{ui}$ and $\lambda_{\mathbf{s}_t\mathbf{s}_{t-1}}$ are associated with the characterizing functions $b_i(\mathbf{o}_u,\mathbf{s}_t)$ and $a(\mathbf{s}_t,\mathbf{s}_{t-1})$.
- $w = 2c+1$ is the number of frames in the acoustic context window
- $Z_\Lambda(\mathbf{O})$ (Zustandsumme) is a normalization coefficient referred to as the partition function, following terminology originally used in statistical mechanics, and often applied to undirected graphical models.

Equation (8) combines different state scores (i.e., $\sum_{u=t-c}^{t+c}\sum_{i=1}^{d^{\mathrm{Aug}}}\lambda_{\mathbf{s}_t}^{ui}b_i(\mathbf{o}_u,\mathbf{s}_t)$) of different frames u and is related to the HMM emission probability score. The state characterizing function $b_i(\mathbf{o}_t,\mathbf{s}_t)$ for each frame was given in (3); $a(\mathbf{s}_t,\mathbf{s}_{t-1})$ is binary valued and can be used to specify the transition topology. ACRF linear decision boundaries provided by (8), are constructed using the first-order statistics accumulated from the augmented space observations. In low-dimensional HMMs, second-order statistics are used; in this case we assume that first order statistics are sufficient, owing to the increased likelihood of linear separability in high-dimensional spaces.

HMMs and ACRFs (in general, linear chain CRFs) share the first-order Markov assumption, which simplifies the training and decoding algorithms. However, unlike HMMs, ACRFs do not assume observation independence and causality, as the joint event in this case is factorized as a simple product of exponential functions. Therefore, the observations and the characterizing functions can be statistically dependent or correlated and can depend on the past and future acoustic context. As a result, ACRFs provide a principled way to relax the HMM conditional independence assumption. The partition function $Z_\Lambda(\mathbf{O})$ is given by

$$Z_\Lambda(\mathbf{O}) = \sum_{\mathbf{S}}\prod_{t=1}^{T}\exp\left(\lambda_{\mathbf{s}_t\mathbf{s}_{t-1}}a(\mathbf{s}_t,\mathbf{s}_{t-1}) + \sum_{u=t-c}^{t+c}\sum_{i=1}^{d^{\mathrm{Aug}}}\lambda_{\mathbf{s}_t}^{ui}b_i(\mathbf{o}_u,\mathbf{s}_t)\right) \quad (9)$$

and it is similar to the total probability $p(\mathbf{O}\,|\,\mathcal{M})$ in HMMs, which can be calculated using the forward algorithm [40].

The ACRF model takes advantage of the construction of augmented spaces to model the acoustic context. It may be expected that modeling acoustic context in augmented spaces within the ACRF framework is an effective technique since the augmented space confusability is expected to be less than for low-dimensional spaces. This additional context may increase discrimination within the acoustic modeling process.

IV. ACRF Optimization

Despite ACRFs having fewer assumptions than HMMs, neither ACRFs nor HMMs are exact models of speech generation. Moreover, without infinite training data, Nadas' conditions for generative modeling [6] are not met, and discriminative training may lead to reduced error rates. ACRF discriminative training takes a similar form to HMM discriminative training [7]–[10], [12] differing only in the constraint functions and the update equations.

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

$$\begin{aligned}\mathcal{F}_{\mathrm{CML}}(\Lambda) &= \sum_{r=1}^{R}\log P_\Lambda(\mathcal{M}_{W_r}\,|\,\mathbf{O}_r)\\ &= \sum_{r=1}^{R}\log\frac{P(W_r)\sum_{\mathbf{S}\,|\,W_r}\exp\sum_t^T\Psi(\mathbf{O},\mathbf{S},c,\Lambda)}{\sum_{\hat{W}}P(\hat{W})\sum_{\mathbf{S}|\hat{W}}\exp\sum_t^T\Psi(\mathbf{O},\mathbf{S},c,\Lambda)}\\ &\approx \sum_{r=1}^{R}\log Z_\Lambda(\mathbf{O}_r\,|\,\mathcal{M}^{\mathrm{num}}) - \log Z_\Lambda(\mathbf{O}_r\,|\,\mathcal{M}^{\mathrm{den}})\end{aligned} \quad (10)$$

where

$$\Psi(\mathbf{O},\mathbf{S},c,\Lambda) = \lambda_{\mathbf{s}_t\mathbf{s}_{t-1}}a(\mathbf{s}_t,\mathbf{s}_{t-1}) + \sum_{u=t-c}^{t+c}\sum_{i=1}^{d^{\mathrm{Aug}}}\lambda_{\mathbf{s}_t}^{ui}b_i(\mathbf{o}_u,\mathbf{s}_t). \quad (11)$$

The optimal parameters Λ^* are estimated by maximizing the CML criterion, which implies minimizing the cross entropy

Authorized licensed use limited to: The University of Edinburgh. Downloaded on February 10, 2009 at 12:51 from IEEE Xplore. Restrictions apply.

between the correct transcription model and the hypothesized recognition model. In other words, the process maximizes the partition function of the correct models[4] (the numerator term) $Z_\Lambda(\mathbf{O}_r \,|\, \mathcal{M}^{\text{num}})$, and simultaneously minimizes the partition function of the recognition model (the denominator term) $Z_\Lambda(\mathbf{O}_r \,|\, \mathcal{M}^{\text{den}})$. The optimal parameters are obtained when the gradient of the CML criterion is zero.

A. Numerical Optimization for ACRFs

Two hill-climbing methods were introduced by Lafferty *et al.* to estimate the parameters of linear chain or sequential CRFs based on the iterative scaling algorithm [40]. The two methods ensure stable update of the objective function, but their speed is a function of the sequence length. Hence, they are very slow and impractical for large tasks. To enable faster training, sequential CRFs are often trained using gradient-based approaches. These methods rely on a locally linear or quadratic approximation by expanding the CML nonlinear objective function $\mathcal{F}_{\text{CML}}(\Lambda+\delta)$ using Taylor's expansion around the current model point Λ in the parameter space [42]. Such approaches are well established in artificial neural networks research [43], [28]. For example, the CRF training process has been accelerated by using a stochastic meta-descent algorithm which utilizes second-order information to adapt the gradient step sizes [44]. Similar methods have been used to train HMMs by relaxing the probabilistic constraints during the HMM training process [45].

For an e-family activation function based on first-order sufficient statistics, the gradient of the CML objective function is given by

$$\nabla \mathcal{F}_{\text{CML}}(\mathbf{O}) = \mathcal{C}_{ji}^{\text{num}}(\mathbf{O}) - \mathcal{C}_{ji}^{\text{den}}(\mathbf{O}) \tag{12}$$

where the sparse accumulators of the sufficient statistics $\mathcal{C}_{ji}(\mathbf{O})$ for the jth state and ith constraint are calculated as follows:

$$\mathcal{C}_{ji}^{\text{num}}(\mathbf{O}) = \sum_{r=1}^{R}\sum_{t=1}^{T_r} \gamma_j^r(t \,|\, \mathcal{M}^{\text{num}}) \mathbf{o}_{rti}^{\text{Aug}} \tag{13}$$

$$\mathcal{C}_{ji}^{\text{den}}(\mathbf{O}) = \sum_{r=1}^{R}\sum_{t=1}^{T_r} \gamma_j^r(t \,|\, \mathcal{M}^{\text{den}}) \mathbf{o}_{rti}^{\text{Aug}} \tag{14}$$

where r is the utterance index and the frame-state alignment probability γ_j, denoting the probability of being in state j at some time t can be written in terms of the forward score $\alpha_j(t)$ and the backward score $\beta_j(t)$ as in HMMs

$$\begin{aligned}\gamma_j(t \,|\, \mathcal{M}) &= P(\mathbf{s}_t = j \,|\, \mathbf{O}; \mathcal{M}) \\ &= \frac{\alpha_j(t \,|\, \mathcal{M})\beta_j(t \,|\, \mathcal{M})}{Z_\Lambda(\mathbf{O} \,|\, \mathcal{M})}.\end{aligned} \tag{15}$$

As discussed in Section VII-B, we did not use a gradient-based method to train ACRFs, although the gradient is needed in the discriminant compression algorithm used to prune the parameter space (Section V-A).

[4]Since a summation over potential functions is commonly called the partition function in undirected graphical modeling, we coin the notation $Z_\Lambda(\mathbf{O}_r \,|\, \mathcal{M}^{\text{num}})$ for the summation of all possible state sequences of the correct models.

B. Approximate Iterative Scaling for ACRFs

Batch lower bound optimization is used to train ACRFs since these methods usually tend to be less heuristic than numerical optimization methods. Exact lower bound optimization of linear chain CRFs [40] based on iterative scaling (IS) variants [46], [47] is very slow. In our work, we use a family of iterative scaling algorithms, which we call *Approximate Iterative Scaling* (AIS), to speed up the training process. While AIS algorithms follow the exact lower bound derivation, they use less conservative steps to provide fast training. Therefore, they do not guarantee the increase of the CML objective function at each iteration, but empirically we have observed few problems with the stability of the algorithm. The AIS algorithms integrate prior knowledge related to the problem formulation (i.e., the observation space construction and the model structure) into the training process. Moreover, the AIS algorithms may increase the rate of convergence by assigning different learning rates for each state or for individual Lagrange multipliers using a learning rate adaptation algorithm [48].

An AIS algorithm updates an e-family activation function based on first-order sufficient statistics, as shown in the following equation:

$$\lambda_{ji}^{\tau+1}(\mathbf{O}) = \lambda_{ji}^{\tau}(\mathbf{O}) + \eta_{\text{AIS}} \log \frac{\mathcal{C}_{ji}^{\text{num}}(\mathbf{O})}{\mathcal{C}_{ji}^{\text{den}}(\mathbf{O})} \tag{16}$$

where η_{AIS} is called the *learning rate* and τ is the iteration number. If the value of the learning rate is sufficiently large, faster training is expected, but there is no guarantee that the algorithm will converge. When the learning rate is extremely small, the updates guarantee an increase in the objective function. The tuning of the learning rate is task dependent with the choice of learning rate balancing word error rate reduction and the number of iterations required for training.

A good heuristic for choosing a suitable initial value for the learning rate is given by

$$\eta_{\text{AIS}} = \frac{G}{w} \tag{17}$$

where $w = 2c + 1$ is the number of frames of acoustic context and G is a global learning rate, which has a default value of $G = 1$. This heuristic is usually sufficient to ensure the increase of the CML objective function and provide fast training of ACRFs. The parameters associated with the state space constraints are not updated and kept fixed during training similar to the language model parameters. This is related to experience, which suggests that updating transition probabilities in HMMs does not lead to an improvement in speech recognition accuracy. Updating the state space constraints using the proposed value of the learning rate leads to instability in the training algorithm.

An AIS algorithm is computationally efficient provided the following conditions are met.

1) Constraints must take positive values in order to minimize the storage complexity of the sparse accumulators. This leads to the use of e-family parametric constraints (see Section II). The size of the accumulators for the constraints which can have negative values will be doubled compared to constraints which have positive values [48].

Authorized licensed use limited to: The University of Edinburgh. Downloaded on February 10, 2009 at 12:51 from IEEE Xplore. Restrictions apply.

2) Binary constraints usually result in very small η_{AIS} values and lead to slow training. Therefore, we do not use any binary constraints in the augmented space formulation. On the other hand, binary constraints associated with the state transition characterizing functions and the language transition characterizing functions are not updated. The omission of these feature functions allows us to use a larger η_{AIS} to update the acoustic constraints, which are directly related to the discrimination between speech classes.
3) The problem is formulated such that $M(\mathbf{o}_t^{\text{Aug}}) = \sum_i \mathbf{o}_{ti}^{\text{Aug}}$ is constant for each observation. In this case it can be shown that the update equations of the improved iterative scaling (IIS) algorithm [47] have the same form as (16), enabling an efficient update without requiring inner loops over the training data. In this paper, we set $M(\mathbf{o}_t^{\text{Aug}}) = w$.
4) The formulation of the constraints should satisfy the conditions arising from the derivation of the iterative scaling (IS) algorithm [47], which uses Jensen's inequality to establish a lower bound on the CML criterion in order to decouple all parameters. The use of Jensen's inequality assumes that the values of the constraints are represented as a posterior probability over a discrete random variable. As a result, constraint formulation based on (3) is desirable to match this derivation. Consequently, the normalization step, although not theoretically necessary, is useful in order to reduce the number of IS iterations for training.

These four conditions were imposed on the augmented space formulation and the parameter update process, resulting in an efficient ACRF training process based on a few iterations of the batch AIS algorithm, outlined above.

V. ACRF Discriminant Compression

ACRF models associated with augmented spaces will lead to a complex training problem with a very large number of parameters. This large number of parameters corresponds to an augmented matrix of size $|\mathbf{s}| \times wd^{\text{Aug}}$, where $|\mathbf{s}|$ represents the total number of states in the system and wd^{Aug} is the dimensionality of the constructed augmented space.[5] Training a large number of parameters can lead to overfitting and poor generalization due to the curse of dimensionality [49]. To address this, we have employed an l_1 regularizer (Section V-A), which we use in the context of an efficient, incremental training algorithm (Section V-B).

A. l_1-ACRF Models

Regularization is a common approach to overcome poor generalization and to provide effective complexity control. In this paper, regularization is achieved by adding an l_1 norm penalty term to the CML criterion as shown in (18)

$$\mathcal{F}_{\text{RCML}}(\Lambda) = \sum_{r=1}^{R} (\log Z_\Lambda(\mathbf{O}_r \,|\, \mathcal{M}^{\text{num}}) - \log Z_\Lambda(\mathbf{O}_r \,|\, \mathcal{M}^{\text{den}})) - \alpha \sum_{s,d^{\text{Aug}}} |\lambda_{ji}|. \quad (18)$$

[5]For example, each phone of TIMIT is represented using a three state ACRF, leading to $48 * 3 = 144$ states in total. An augmented space with dimensionality $d^{\text{Aug}} = 129,295$, will lead to 18, 618, and 480 parameters that must be estimated robustly.

The l_1 regularizer or Lasso penalty $\sum |\lambda_{ji}|$ is often used to increase the sparseness of the model since it can lead to solutions where some elements of λ_{ji} are exactly zero [50].

The gradient of the $\mathcal{F}_{\text{RCML}}$ objective function is given by

$$\frac{\partial \mathcal{F}_{\text{RCML}}}{\partial \lambda_{ji}} = \sum_{r=1}^{R} \left(\mathcal{C}_{ji}^{\text{num}}(\mathbf{O}_r \,|\, \mathcal{M}^{\text{num}}) - \mathcal{C}_{ji}^{\text{den}}(\mathbf{O}_r \,|\, \mathcal{M}^{\text{den}})\right) - \alpha \,\text{sign}(\lambda_{ji}) \quad (19)$$

where the gradient of $\mathcal{F}_{\text{RCML}}$ can be defined for points other than $\lambda_{ji} = 0$ since

$$\frac{\partial |\lambda_{ji}|}{\partial \lambda_{ji}} = \begin{cases} +1, & \text{for } \lambda_{ji} > 0 \\ \text{undefined}, & \text{for } \lambda_{ji} = 0 \\ -1, & \text{for } \lambda_{ji} < 0. \end{cases} \quad (20)$$

$\text{sign}(\lambda_{ji})$ is substituted by -1 or $+1$ at $\lambda_{ji} = 0$, to ensure the increase of the $\mathcal{F}_{\text{RCML}}$ objective function, solving the undefined gradient problem associated with an l_1-norm regularizer.

When $\partial \mathcal{F}_{\text{CML}}/\partial \lambda_{ji} > \alpha$, this means that $\partial \mathcal{F}_{\text{RCML}}/\partial \lambda_{ji} > 0$ regardless of the sign of λ_{ji}. Since λ_{ji} is zero and $\text{sign}(\lambda_{ji})$ is not defined, a choice of $\text{sign}(\lambda_{ji}) = +1$ can increase $\mathcal{F}_{\text{RCML}}$; similarly, if $\partial \mathcal{F}_{\text{CML}}/\partial \lambda_{ji} < -\alpha$, then a choice of $\text{sign}(\lambda_{ji}) = -1$ will increase $\mathcal{F}_{\text{RCML}}$. In short, $|\partial \mathcal{F}_{\text{CML}}/\partial \lambda_{ji}| > \alpha$ specifies an evaluation condition for useful parameter for modeling in the l_1 norm sense. The parameters where $|\partial \mathcal{F}_{\text{CML}}/\partial \lambda_{ji}| < \alpha$, are not included in the model. Hence, the l_1 norm leads to sparse solutions and specifies the maximum number of parameters that can be added to the model $\mathcal{M}$. This evaluation condition and the AIS algorithm form an incremental optimization algorithm to train ACRFs and prune their parameter space concurrently. Alternatively, gradient based optimization can be used to train general models as the gradient is defined and calculated [51].

The value for the hyperparameter α specifies the compromise between the complexity of the model and modeling accuracy. Increasing the value of α will lead to reduction of the number of the active parameters in the model $\mathcal{M}$. Selecting a suitable value for α can usually be achieved via cross validation. Alternatively, the problem can also be cast as model selection within the marginal likelihood or evidence framework [52]. A simple and pragmatic method that can be useful for large-scale optimization required for speech recognition is proposed. This method is based on training the unregularized simple ACRFs ($c = 0$ and $\alpha = 0$) for few iterations and recording the best error rate for heldout/test data. Then, the value of α is increased gradually and the l_1-ACRFs are retrained, selecting a value of α which leads to a significant reduction in the number of parameters with the minimum reduction in the recognition accuracy. Later, α is fixed during the l_1-ACRFs training stage of the incremental context integration ($c > 0$).

The discriminative pruning method described here is practical for large-scale problems such as speech recognition and will lead to very sparse models as the results show in Section VI. Other pruning methods have been developed based on forward greedy constraint induction [47], Optimal Brain Damage [53], or Optimal Brain Surgeon [54]. However, these methods scale

Authorized licensed use limited to: The University of Edinburgh. Downloaded on February 10, 2009 at 12:51 from IEEE Xplore. Restrictions apply.

poorly to speech recognition.[6] Some researchers prune the parameter space by removing the parameters associated with constraints that have low empirical expectation values before the training process[7] [56], [57]. This technique belongs to *learning model parameters* methods only, while fixing the structure of the model. Of course, *learning model structure* is a harder problem with respect to learning the parameters of a specified model only. Our method learns the model structure and the parameters concurrently in an efficient way.

B. Incremental AIS Algorithm for l_1-ACRFs

l_1-ACRFs were trained using the incremental training algorithm detailed in algorithm 1. The algorithm starts with empty model $\mathcal{M} = \phi$. At each iteration, the evaluation condition $|\partial \mathcal{F}_{\mathrm{CML}}/\partial \lambda_{ji}| > \alpha$ is calculated for each parameter, which is not part of the current model. If any parameter is able to pass this gradient test, we add it to the model $\mathcal{M} \cup \lambda_{ji}(\mathbf{O})$. All the parameters which are elements of the current model $\lambda_{ji}(\mathbf{O}) \in \mathcal{M}$ are updated using AIS step. Note that this algorithm 1 has no inner loops over the training data. Finally, the process stop according to a valid termination condition.

Algorithm 1: Incremental AIS Algorithm for l_1-ACRFs

Input: $\mathcal{M} = \phi, \eta_{\mathrm{AIS}} = (G/w)$ for context modeling $\mathcal{I} \leq 25$ and $\alpha > 0$

repeat

Accumulate $\mathcal{C}_{ji}^{\mathrm{num}}(\mathbf{O}_r \,|\, \mathcal{M}^{\mathrm{num}})$ and $\mathcal{C}_{ji}^{\mathrm{den}}(\mathbf{O}_r \,|\, \mathcal{M}^{\mathrm{den}}), \forall r$;

forall $\lambda_{ji}(\mathbf{O}) \notin \mathcal{M}$ **do**//*select parameters*

if $|\mathcal{C}_{ji}^{\mathrm{num}}(\mathbf{O} \,|\, \mathcal{M}^{\mathrm{num}}) - \mathcal{C}_{ji}^{\mathrm{den}}(\mathbf{O} \,|\, \mathcal{M}^{\mathrm{den}})| > \alpha$ **then**

$\mathcal{M} \cup \lambda_{ji}(\mathbf{O})$;

end

forall $\lambda_{ji}(\mathbf{O}) \in \mathcal{M}$ **do**//*update parameters*

$\lambda_{ji}^{t+1}(\mathbf{O}) = \lambda_{ji}^{t}(\mathbf{O}) + \eta_{\mathrm{AIS}} \log(\mathcal{C}_{ji}^{\mathrm{num}}(\mathbf{O} \,|\, \mathcal{M}^{\mathrm{num}}))/(\mathcal{C}_{ji}^{\mathrm{den}}(\mathbf{O} \,|\, \mathcal{M}^{\mathrm{den}}))$;

if $c > 0$ **then** adapt η_{AIS}//*optional*

end

until $\mathrm{itr} < \mathcal{I}$.

In the case of context modeling, an adaptive version of the AIS algorithm can be used to accelerate the training process. In general, the actual c value, which leads to the optimal recognition performance is task dependent. As a result, it should be optimized during the training as a hyperparameter. This may be done using incremental training by adding two context frames (left and right) and retraining the whole system. Therefore, the process of adding two context frames incrementally forces a parsimony strategy to find the optimal c value. Changing the value of c during the incremental context integration directly changes the dimensionality of the augmented space d^{Aug}, and the decoding parameters should be optimized for each value of c for the same task.

VI. Experiments

We have carried out phone recognition experiments on the TIMIT corpus.[8] We used the 462 speaker training set, testing primarily on the 168 speaker full test set; we also report results using the 24-speaker core test set. Unless otherwise indicated, the SA1 and SA2 utterances were not used. The speech was analyzed using a 25-ms Hamming window with a 10-ms fixed frame rate. In all the experiments we represented the speech using 12th-order Mel frequency cepstral coefficients (MFCCs), energy, along with their first and second temporal derivatives, resulting in a 39-element feature vector. Following Lee [58], the original 61 phone classes in TIMIT were mapped to a set of 48 labels, which were used for training. This set of 48 phone classes was mapped down to a set of 39 classes [58], after decoding, and out phone recognition results are reported on these classes, in terms of the phone error rate (PER), which is analogous to word error rate. All our experiments used a bigram language model over phones, estimated from the training set.

As a baseline we trained both context-independent (CI) and context-dependent (CD) GMM/HMM phone recognizers, using HTK[9] [59]. The CI system contained 144 emitting states, with 55 mixture components per state; the CD system contained 1127 physical states, with 20 mixture components per state. On the full test set, the CI system resulted in 29.2% PER, and the CD system resulted in 27.3% PER. On the core test set, the PERs were 30.1% (CI) and 27.3% (CD). Note that the CD system results from augmenting the state space (see Section I) and thus should not be compared directly with the CI l_1-ACRFs described below. The HMM systems all operate in the standard 39-dimensional feature space; the high-dimensional augmented space requires dimension reduction (such as the discriminant compression that we employ when estimating the l_1-ACRF models). In principle, a technique such as linear discriminant analysis could be used to project the high-dimensional representations to low dimensions, but this is algorithmically complex because of the size of the required sparse covariance matrices. On the other hand, projection methods such as the fMPE approach [21] can be used to provide a low-dimensional representation that can be suitable for HMMs (Section VII). However, such projection methods do not address the acoustic modeling formulation and are mainly used to overcome those HMM limitations that we aim to relax explicitly (Section I).

When training the ACRF models, the acoustic space was partitioned into about 7000 Gaussian (e-family) activation functions, calculated in advance for each utterance in the database.

[6]Optimal Brain Damage and Optimal Brain Surgeon methods are based on building large models and using the Hessian matrices of these large models to compute a saliency measure for parameters and make backward model selection. However, building large models and then pruning them is impractical speech recognition systems. On the other hand, it was shown that constraint induction based on mean field approximation is only efficient for binary constraints [47]. For continuous spaces, this method needed inner loops over the training data [55].

[7]Theoretically, we should remove the parameters that do not improve the CML objective function.

[8]http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

[9]http://htk.eng.cam.ac.uk/

Authorized licensed use limited to: The University of Edinburgh. Downloaded on February 10, 2009 at 12:51 from IEEE Xplore. Restrictions apply.

TABLE I
UNREGULARIZED ACRF DECODING RESULTS ($c = 0$ AND $\alpha = 0$) WITH THE TIMIT FULL TEST SET

#Itr	Corr	Sub	Del	Ins	PER
1	70.4	20.1	9.5	3.3	33.0%
2	72.1	19.5	8.4	3.8	31.7%
3	72.6	19.4	8.0	3.9	31.3%
4	72.8	19.3	7.9	4.0	31.2%
5	72.9	19.3	7.8	4.1	31.2%

as described in Section II. Each phone was represented using a three state left-to-right ACRF, all parameters related to the constraint functions were initialized to zero and the transition probabilities were initialized either from trained HMM models or from a uniform transition matrix, forcing left to right ACRFs. The training procedure accumulated the $\mathcal{M}^{\text{num}}$ sufficient statistics via a Viterbi pass (forced alignment) of the reference transcription using HMMs trained using maximum likelihood and the $\mathcal{M}^{\text{den}}$ sufficient statistics were approximated with state estimates, to avoid the necessity of building lattices in the context modeling stage [60]. The models were trained using the Incremental AIS algorithm described in Section V-B.

In Table I, we report phone recognition results, for the full test set, using unregularized ACRFs with no acoustic context ($c = 0$). All parameters of the ACRF models were initialized from zero, with the optimization reaching a good solution after a few iterations. This fast training speed arises since the augmented spaces are normalized and the features of the augmented spaces are approximately uncorrelated (i.e., the relation between features is a soft winner-take-all relationship).[10] In this case, AIS training can update all the parameters in the model with a learning rate of $\eta_{\text{AIS}} = 1.0$.

To evaluate the efficiency of the incremental AIS algorithm for l_1-ACRFs, the PERs for different values of the hyperparameter α is given in Table II, again using the full TIMIT test set. Training was stopped after five iterations. We measured the sparseness of the models using a criterion referred to as the compression ratio (CR) of the parameter space, given by

$$\text{CR}(\alpha) = \frac{\#\text{Param}(\alpha = 0.0) - \#\text{Param}(\alpha)}{\#\text{Param}(\alpha = 0.0)}. \qquad (21)$$

Increasing the value of α increases the sparseness of the model. It can be seen that setting $\alpha = 20$ resulted in a compression ratio of 0.97, with a PER increase of 2.2% absolute or 7.0% relative. Since a frame can activate only a specific number of regions or partitions in the acoustic space—those regions related to confusable speech sounds—high compression ratios should indeed be expected with only a limited increase in the error rate. In the following experiments, we set $\alpha = 2$, resulting in a compression ratio of 0.81.

The above experiments, in which an effective dimensionality $d^{\text{Eff}} = 10$ was obtained by taking an n-best shortlist (Section II), may be considered as a "soft" vector quantization (VQ). A "hard" VQ corresponds to taking $d^{\text{Eff}} = 1$. In this case, the PER (on the full test set) is increased to 33.6% ($\alpha = 2.0$).

The usual HMM conditional independence assumptions may be relaxed by employing context modeling, using a temporal window of $\pm c$ frames of context, as discussed in Sections II-B and III. The phone recognition results on the full TIMIT test set, for different amounts of acoustic context are summarized in Table III. For $c = 1$ and $c = 3$, the language model scaling factor was set to 2.0. For $3 < c < 9$, the language model scaling factor was set to 1.0 as the dynamic range of augmented space constraints and language model probabilities are similar. When $c = 9$, it was set to 0.5. Clearly, it can be seen that considering long acoustic context can lead to significant improvement in PER over the baseline system where $c = 0$. On the core test set, using a context of $c = 9$ results in a PER of 26.6%, compared with 32.3% when $c = 0$.

[10]Context modeling may lead to correlation between some features in the augmented spaces.

TABLE II
TIMIT FULL TEST PERS FOR l_1-ACRFs TRAINED WITH DIFFERENT SETTINGS OF HYPERPARAMETER $\alpha(c = 0)$

α	$\lvert\Lambda(\alpha)\rvert$	CR(α)	Corr	Sub	Del	Ins	PER
0.0	324486	0	72.9	19.3	7.8	4.1	31.2%
0.1	323348	≈ 0	72.9	19.3	7.8	4.1	31.2%
0.5	266721	0.18	72.9	19.3	7.8	4.1	31.2%
1.0	139576	0.57	72.7	19.3	8.0	4.0	31.3%
2.0	59943	0.81	72.6	19.4	8.0	4.2	31.6%
5.0	31568	0.90	72.4	19.8	7.8	4.5	32.1%
10.0	18812	0.94	72.5	19.8	7.7	5.0	32.5%
20.0	10609	0.97	72.3	20.0	7.7	5.7	33.4%

TABLE III
PERS ON THE TIMIT FULL TEST SET FOR DIFFERENT AMOUNTS OF CONTEXT MODELING WITHIN THE l_1-ACRF FRAMEWORK ($\alpha = 2.0$)

c	d^{Aug}	d^{Eff}	$\lvert\Lambda(\alpha)\rvert$	Corr	Sub	Del	Ins	PER
0	6780	10	59943	72.6	19.4	8.0	4.2	31.6%
1	20340	30	180893	74.4	18.4	7.2	4.2	29.8%
3	47460	70	457422	74.8	17.1	8.0	3.1	28.3%
5	74580	110	785214	76.7	16.4	7.0	3.5	26.8%
7	101700	150	1154007	76.8	15.7	7.5	2.9	26.1%
9	128820	190	1548672	77.3	15.4	7.3	3.0	25.7%

Explicit acoustic context modeling may increase discrimination within the acoustic modeling process since computing state scores over longer time intervals may reduce the acoustic confusability (analogous to phone spectral properties being more clear over longer time intervals in human spectrogram reading). Since confusability in high-dimensional augmented spaces is less than in low-dimensional spaces, explicit modeling of the acoustic context may be more effective in such settings. In the case of l_1-ACRFs, integrating the acoustic context into the augmented spaces gave $\approx$6% absolute reduction in PER, compared without using acoustic context ($c = 0$). The results in Table III may illuminate why acoustic context can help to improve discrimination. These results show that most improvements in the PER are due to a reduction in substitution errors. This result may suggest that using acoustic context information to compute acoustic scores may be understood as smoothing the state scores as trajectories over longer time intervals. Hence, acoustic context may prevent abrupt jumps in the state space due to short time signal analysis limitations and strong confusability at frame level between similar speech classes.

Table IV compares our results with those reported by others on TIMIT phone recognition. Three different test sets have been used for TIMIT (core, full, and full including SA utterances),

Authorized licensed use limited to: The University of Edinburgh. Downloaded on February 10, 2009 at 12:51 from IEEE Xplore. Restrictions apply.

TABLE IV
COMPARISON BETWEEN DIFFERENT APPROACHES FOR TIMIT PHONE RECOGNITION TASK

Method	Test set	PER
CD-HMMs baseline (HTK)	Full	27.3%
CD-HMMs baseline (HTK)	Core	27.3%
Bayesian Triphone HMM [61]	Full+SA	24.4%
Nonlinear symplectic transformation [67]	Full+SA	24.4%
Monophone l_1-ACRFs (complete train set)	Full+SA	**23.0**%
TRAP (one band expert) [64]	Full	28.2%
Monophone l_1-ACRFs	Full	**25.7**%
Recurrent NN (RNN/HMM) [68]	Full	25.0%
TRAP tandom [64]	Full	21.4%
Knowledge based rescoring [65]	Full	21.0%
HMM CML estimation [63]	Core	31.5%
HMM large margin estimation [63]	Core	28.2%
Monophone l_1-ACRFs	Core	**26.6**%
Recurrent NN (RNN/HMM) [68]	Core	26.1%
Bayesian Triphone HMM [61]	Core	25.6%
Heterogeneous measurements [66]	Core	24.4%

and we report results in all of these cases. Approaches based on augmenting the state space such as CD-HMMs [27] and the Bayesian Triphone HMM [61] address one of the enhancements over the HMM basic formulation (Section I), which is not addressed in the ACRF framework. The other two enhancements are addressed within the ACRF framework. In general, improvements based on using different objective functions [62], [63] do not address the acoustic modeling formulation and can be used to train l_1-ACRFs as well. l_1-ACRFs can also take advantage of the TRAP tandem approach [64], [65] as a powerful frontend [20]. System combination and rescoring, committee based methods, such as [65], [66] may be applied for l_1-ACRFs. Note that 50 speakers from the full test set was used for cross validation in [66].

Bilmes [32] experimentally compared acoustic context modeling with the basic HMM formulation, to measure the gain from relaxing the conditional independence assumptions. Making such an experimental comparison from the ACRF experiments, results in an improvement in PER of about 6%, which is consistent across the different test sets that we investigated. This result suggests that the ACRFs learned intrinsic information related to the context rather than being tuned to a specific test set.

VII. DISCUSSION

In this section, we discuss a number of the issues that arise from the l_1-ACRF framework, including scalability, optimization, and the relationship of this approach to the conventional HMM framework and to a number of recently proposed discriminative approaches to acoustic modeling.

A. Scalability

The theoretical number of additional parameters to be estimated within the l_1-ACRF framework is $|\Lambda(\alpha)| = (|\mathbf{s}| \times |\mathbf{g}| \times w)|_{\alpha=0}$, where $|\mathbf{s}|$ is the number of states in the system, $|\mathbf{g}|$ is the number of Gaussians in the system, w is the width of the context window, and α is the compression hyperparameter. The number of parameters associated with the estimated Gaussians are shared between HMM and l_1-ACRF frameworks and they are not counted in $|\Lambda(\alpha)|$. This may result in billions of parameters, so the technique may not scale directly. However, we have shown that if $\alpha \gg 0$, l_1-ACRFs may achieve high compression ratios. If the context information is integrated incrementally (one additional frame of left and right context at a time), then this can force parsimony while discovering useful context dependencies. Furthermore, context compression is done concurrently with AIS training and does not lead to inner loops over the training data. In general, high compression ratios may be expected since a frame can activate only a specific number of regions or partitions in the acoustic space. On the other hand, the number of Gaussians ($|\mathbf{g}|$) can be controlled by reclustering the original set of Gaussians in an acoustic model into a smaller set, as done in the fMPE framework [21], [22].

B. Optimization

One of the principal advantages of HMM-based acoustic modeling is the availability of the EM algorithm for parameter estimation, which does not require the tuning of hyperparameters. In the case of ACRFs, an EM algorithm is not available, and the available optimization techniques for this class of model are based on either iterative scaling or gradient descent.

In the case of maximum entropy modeling for natural language processing, Malouf [69] demonstrated that gradient-based optimization is considerably more efficient than approaches based on iterative scaling. However, we note that in this case the structure of the problem—binary features and highly correlated constraints (features)—was ill-matched to iterative scaling. The drawback of gradient-based optimization is the necessity of estimating hyperparamers, such as the learning rate. In our case, where the model changes during the training procedure (due to incrementing the acoustic context) hyperparameter estimation becomes demanding.

Stochastic or online updates based on gradient descent algorithms have proven to be very efficient for a number of large-scale problems [70]. In our work, we investigated the usage of online iterative scaling algorithm [48]. However, while it was easy to show it provides faster convergence, we found this advantage outweighed by the necessity of hyperparameter estimation, which included step sizes, block sizes, and smoothing parameters, and difficulties in implementation in a batch submission parallel environment. Furthermore, the compression method that we employ requires gradient estimation, and is less reliable if online (stochastic) gradient estimates are used.

C. Related Work

l_1-ACRF acoustic modeling builds on conventional HMM approaches. For example, the augmented space may be constructed using the Gaussian components of a GMM/HMM acoustic model, and an HMM augmented state space may be used to initialize an l_1-ACRF state space without the need for clustering. Without acoustic context ($c = 0$), a discriminatively trained HMM and the l_1-ACRF frameworks are similar, differing only in the constraint functions—i.e., the use of pruned posterior probability estimates (3) in l_1-ACRFs—and the optimization process. There are also close links to some recently introduced techniques for acoustic modeling, including maximum entropy approaches, discriminative feature projections,

Authorized licensed use limited to: The University of Edinburgh. Downloaded on February 10, 2009 at 12:51 from IEEE Xplore. Restrictions apply.

and various approaches that attempt to relax the conditional independence assumptions.

Rank-based scoring used in maximum entropy direct modeling approaches [57], [71] may be interpreted as a means to construct an augmented space. In this case, the direct model approach is similar to the l_1-ACRF framework when $c = 0$ and does not take advantage of acoustic context information.

Buried Markov models (BMMs) are directed graphical models which extend the HMM by integrating specific cross-observation dependencies to relax HMM conditional independence assumptions [31], [32]. The BMM approach may be used for data driven sparse acoustic context modeling in the original (low-dimensional) feature space within the HMM framework. Thus, l_1-ACRFs and BMMs are conceptually similar. However, the major difference between BMMs and l_1-ACRFs is that l_1-ACRFs learn the sparse structure of the model and the parameters concurrently and the compression of acoustic context information is done in the augmented space rather than the original feature space. Moreover, in the l_1-ACRF framework contextual information is integrated incrementally to maintain a degree of parsimony controlled by the recognition performance.

The fMPE framework [21] is a nonlinear discriminant feature projection method, which was motivated and developed within the HMM framework. The fMPE framework estimates an augmented projection matrix M, which is used to construct a new feature vector (i.e., $\mathbf{o}_t^{\text{new}} = \mathbf{o}_t + M\mathbf{o}_t^{\text{Aug}}$). Thus, fMPE estimates a correction factor based on a projection method to improve the discrimination between speech classes without changing the system's dimensionality (i.e., $\mathbf{o}_t^{\text{new}}$ and $\mathbf{o}_t$ have the same dimensionality). The augmented spaces used in the l_1-ACRF framework are similar to the augmented spaces used in the fMPE framework. In an improved fMPE formulation [22], the original set of Gaussians in an acoustic model is reclustered into a smaller set of Gaussians which simplifies the construction of an augmented space with a dimensionality of approximately 10^3. In general, it is expected that the number of parameters in the l_1-ACRF framework will be huge with respect to the number of parameters in the fMPE framework (hence, it is not necessary to compress the projection matrix in fMPE). fMPE feature projection methods may be used within l_1-ACRF framework.

Maximum entropy acoustic modeling based on low-dimensional spaces has became an active area of research [72]–[74], [48]. A linear chain CRF model analogous to an HMM (as it used in speech recognition) relaxes the stochastic transition constraints and its local observation scoring is based on quadratic activation functions (6). For instance, hidden conditional random fields (HCRFs) [74], which are linear chain CRFs graphical models, are formulated based on low-dimensional spaces similar to HMMs. The graphical model behind HCRFs has identical conditional independence properties to HMMs, but the HCRF approach trains the acoustic and language constraints in a unified model. The activation functions in (6) are more flexible discriminant functions than Gaussian densities, which are used for local observation scoring within the HMM (but the physical meaning of mean and variance is no longer available). The goal of the training process is to estimate the weights of these e-family activation functions as well as the parameters associated with transition constraints.

Score-space kernels [75], [76], which are a generalization of the Fisher kernel [77], are used to extract new sufficient statistics, which may relax the conditional independence assumptions in a systematic fashion. These sufficient statistics are used to train conditional statistical models (C-Aug) for postprocessing in HMM-based speech recognition [78]. The form of the augmented models is

$$P_\Lambda(W \mid \mathbf{O}) = \frac{1}{Z_\Lambda(\mathbf{O})} p_o(\mathbf{O} \mid W) \exp\left(\lambda_W^T f(W, \mathbf{O}, \Lambda)\right) \tag{22}$$

where $p_o(\mathbf{O} \mid W)$ represents our prior knowledge (often an HMM) as a reference distribution and $f(W, \mathbf{O}, \Lambda)$ are additional constraints (the new sufficient statistics) provided by score-space kernels. Layton and Gales [78] relate $f(W, \mathbf{O}, \Lambda)$ and $p_o(\mathbf{O} \mid W)$ using a local exponential expansion but they can be independent. Thus, the C-Aug framework provides an augmentation mechanism which is different from the l_1-ACRF framework, which augments the acoustic model by imposing a huge number of constraints in a sparse high-dimensional space. However, both l_1-ACRF and C-Aug may be interpreted as undirected graphical models, related to maximum entropy approaches. We note that it would be possible to apply C-Aug modeling as a postprocessing step in the l_1-ACRF framework.

VIII. Summary

The basic mathematical theory and an efficient implementation of l_1-ACRF acoustic modeling have been presented. Within the l_1-ACRF framework, the use of high-dimensional spaces to reduce confusability and the use of acoustic context information to handle the sequential phenomena of the speech signal lead to sparse context modeling in an augmented space, which lead to improved discrimination and lower error rates on the TIMIT phone recognition task. Frame-based acoustic models based on l_1-ACRFs and HMMs have some similarities; in particular, both approaches have similar training speed and decoding algorithms. Hence, l_1-ACRF acoustic modeling attempts to address some of the limitations of HMMs while maintaining many of the good aspects, which have made them successful.

References

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[2] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press, 1997.

[3] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Upper Saddle River, NJ: Prentice-Hall, 2001.

[4] J. Bilmes, "What HMMs can do," *IEICE Trans. Inf. Syst.*, vol. E89-D, no. 3, pp. 869–891, 2006.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[6] A. Nadas, "A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP–31, no. 4, pp. 814–817, Aug. 1983.

Authorized licensed use limited to: The University of Edinburgh. Downloaded on February 10, 2009 at 12:51 from IEEE Xplore. Restrictions apply.

[7] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. IEEE ICASSP*, Tokyo, Japan, 1986, pp. 49–52.

[8] P. F. Brown, "The acoustic-modelling problem in automatic speech recognition," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, 1987.

[9] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo, "An inequality for rational function with applications to some statistical estimation problems," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 107–113, Jan. 1991.

[10] Y. Normandin, "Hidden Markov Models, Maximum Mutual Information Estimation and the Speech Recognition Problem," Ph.D. dissertation, McGill Univ., Montreal, QC, Canada, 1991.

[11] R. Schlüter, W. Macherey, B. Müller, and H. Ney, "Comparison of discriminative training criteria and optimization methods for speech recognition," *Speech Commun.* vol. 34, no. 3, pp. 287–310, 2001 [Online]. Available: citeseer.ist.psu.edu/article/uter01comparison.html.

[12] P. Woodland and D. Povey, "Large scale discriminative training for speech recognition," in *Proc. ISCA ITRW Automatic Speech Recognition: Challenges for the Millenium*, 2000, pp. 7–16.

[13] S. Young, "A review of large-vocabulary continuous-speech," *IEEE Signal Process. Mag.*, vol. 13, no. 5, pp. 45–57, Sep. 1996.

[14] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary speech recognition," in *Proc. IEEE ICASSP*, San Francisco, CA, 1992, vol. 1, pp. 13–16.

[15] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Commun.*, vol. 26, no. 4, pp. 283–297, 1998.

[16] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Proc. IEEE ICASSP*, Istanbul, Turkey, 2000, vol. 2, pp. 1129–1132 [Online]. Available: citeseer.ist.psu.edu/article/saon00maximum.html.

[17] M. Rahim and C. H. Lee, "Simultaneous ANN feature and HMM recognizer design using string-based minimum classification error (MCE) training," in *Proc. ICSLP*, Philadelphia, PA, 1996, pp. 1824–1827.

[18] H. Hermansky and S. Sharma, "TRAPs—Classifiers of temporal patterns," in *Proc. ICSLP*, 1998, pp. 1003–1006.

[19] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature stream extraction for conventional HMM systems," in *Proc. IEEE ICASSP*, Istanbul, Turkey, 2000, pp. 1635–1638.

[20] N. Morgan, Z. Qifeng, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cretin, H. Bourlard, and M. Athineos, "Pushing the envelope—Aside," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 81–88, Sept. 2005.

[21] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proc. IEEE ICASSP*, Philadelphia, PA, Mar. 2005, vol. 1, pp. 961–964.

[22] D. Povey, "Improvements to fMPE for discriminative training of features," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 2977–2980.

[23] R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul, "Context-dependent modeling for acoustic-phonetic recognition of continuous speech," in *Proc. IEEE ICASSP*, 1985, pp. 1205–1208.

[24] K.-F. Lee, "Large Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, 1988.

[25] M. Hwang and X. Huang, "Acoustic classification of phonetic hidden Markov models," in *Proc. Eurospeech*, Genova, Italy, 1991, pp. 785–788.

[26] L. Bahl, P. deSouza, P. Gopalakrishnan, D. Nahamoo, and M. Picheny, "Decision trees for phonological rules in continuous speech," in *Proc. IEEE ICASSP*, Toronto, ON, Canada, 1991, vol. 1, pp. 185–188.

[27] S. Young and P. Woodland, "State clustering in HMM-based continuous speech recognition," *Comput. Speech, Lang.*, vol. 8, no. 4, pp. 369–384, 1994.

[28] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1998.

[29] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley-Interscience, 1998.

[30] C. J. Wellekens, "Explicit time correlation in hidden Markov models for speech recognition," in *Proc. IEEE ICASSP*, 1987, pp. 384–386.

[31] J. Bilmes, "Data-driven extensions to HMM statistical dependencies," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 69–72.

[32] J. Bilmes, "Buried Markov models for speech recognition," in *Proc. IEEE ICASSP*, Phoenix, AZ, Mar. 1999, vol. 2, pp. 713–716.

[33] S. Furui, "Speaker independent isolated word recognizer using dynamic features of the speech spectrum," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 1, pp. 52–59, Feb. 1986.

[34] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. Electron. Comput.*, vol. EC-14, no. 3, pp. 326–334, Jun. 1965.

[35] B. E. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop Comput. Learning Theory*, 1992, pp. 144–152 [Online]. Available: citeseer.ist.psu.edu/boser92training.html.

[36] E. Bocchieri, "Vector quantization for the efficient computation of continuous density likelihoods," in *Proc. IEEE ICASSP*, Minneapolis, MN, Apr. 1993, vol. 2, pp. 692–694.

[37] M. Gales, K. Knill, and S. Young, "State-based Gaussian selection in large vocabulary continuous speech recognition using HMMs," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 152–161, Mar. 1999.

[38] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Norwell, MA: Kluwer, 1994.

[39] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, and M. A. Picheny, "Robust methods for using context-dependent features and models in a continuous speech recognizer," in *Proc. IEEE ICASSP*, Adelaide, Australia, 1994, pp. I-533–I-536.

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

[41] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proc. IEEE*, vol. 70, no. 9, pp. 939–952, 1982.

[42] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York: Springer, 1999.

[43] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995.

[44] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy, "Accelerated training of conditional random fields with stochastic gradient methods," in *Proc. ICML*, 2006, pp. 969–976.

[45] S. Kapadia, "Discriminative Training of Hidden Markov Models," Ph.D. dissertation, Univ. of Cambridge, Cambridge, U.K., 1998.

[46] J. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *Ann. Math. Statist.*, vol. 43, no. 5, pp. 1470–1480, 1972.

[47] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Inducing features of random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 4, pp. 380–393, Apr. 1997.

[48] Y. Hifny, "Conditional random fields for continuous speech recognition," Ph.D. dissertation, Univ. of Sheffield, Sheffield, U.K., 2006.

[49] R. Bellman, *Adaptive Control Processes*. Princeton, NJ: Princeton Univ. Press, 1961.

[50] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*. New York: Springer, 2001.

[51] S. Perkins, K. Lacker, and J. Theiler, "Grafting: Fast, incremental feature selection by gradient descent in function space," *J. Mach. Learn. Res.*, vol. 3, pp. 1333–1356, 2003.

[52] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. New York: Springer, 1985.

[53] Y. LeCun, J. Denker, and S. Solla, "Optimal brain damage," in *Proc. NIPS*, 1990, vol. 2, pp. 598–605.

[54] B. Hassibi, D. G. Stork, and G. J. Wolff, "Optimal brain surgeon and general network pruning," in *Proc. IEEE Int. Conf. Neural Netw.*, 1993, vol. 1, pp. 293–299.

[55] Y. Hifny, S. Renals, and N. D. Lawrence, "Acoustic space dimensionality selection and combination using the maximum entropy principle," in *Proc. IEEE ICASSP*, Montreal, QC, Canada, May 2004, vol. 5, pp. 637–640.

[56] A. Ratnaparkhi, "A maximum entropy model for part-of-speech tagging," in *Proc. EMNLP*, 1996, pp. 133–142.

[57] A. Likhododev and Y. Gao, "Direct models for phoneme recognition," in *Proc. IEEE ICASSP*, Orlando, FL, May 2002, vol. 1, pp. 89–92.

[58] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

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

[60] Y. Hifny, S. Renals, and N. Lawrence, "A hybrid MaxEnt/HMM based ASR system," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 3017–3020.

[61] J. Ming and F. J. Smith, "Improved phone recognition using Bayesian triphone models," in *Proc. IEEE ICASSP*, Seattle, WA, May 1998, vol. 1, pp. 409–412.

Authorized licensed use limited to: The University of Edinburgh. Downloaded on February 10, 2009 at 12:51 from IEEE Xplore. Restrictions apply.

[62] S. Kapadia, V. Valtchev, and S. Young, "MMI training for continuous phoneme recognition on the TIMIT database," in *Proc. IEEE ICASSP*, Minneapolis, MN, Apr. 1993, vol. 2, pp. 491–494.
[63] F. Sha and L. K. Saul, "Comparison of large margin training to other discriminative methods for phonetic recognition by hidden Markov models," in *Proc. IEEE ICASSP*, Honolulu, HI, 2007, vol. 4, pp. 313–316.
[64] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. IEEE ICASSP*, Toulouse, France, 2006, pp. 325–328.
[65] S. M. Siniscalchi, P. Schwarz, and C.-H. Lee, "High-accuracy phone recognition by combining high-performance lattice generation and knowledge based rescoring," in *Proc. IEEE ICASSP*, Honolulu, HI, 2007, pp. 869–872.
[66] A. Halberstadt and J. Glass, "Heterogeneous measurements and multiple classifiers for speech recognition," in *Proc. ICSLP*, Sydney, Australia, Nov. 1998, vol. 3, pp. 995–998.
[67] M. Omar and M. Hasegawa-Johnson, "Approximately independent factors of speech using nonlinear symplectic transformation," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 660–671, Nov. 2003.
[68] A. Robinson, "An application of recurrent neural nets to phone probability estimation," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 298–305, Mar. 1994.
[69] R. Malouf, "A comparison of algorithms for maximum entropy parameter estimation," in *Proc. CoNLL*, 2002, pp. 49–55.
[70] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Mueller, "Efficient backprop," in *Neural Networks—Tricks of the Trade, Lecture Notes in Computer Sciences 1524*. New York: Springer, 1998, pp. 5–50.
[71] J. K. Hong-Kwang and Y. Gao, "Maximum entropy direct models for speech recognition," in *Proc IEEE ASRU Workshop*, St. Thomas, Virgin Islands, Dec. 2003, pp. 1–6.
[72] K. Van Horn, A maximum-entropy solution to the frame dependency problem in speech recognition Dept. of Comput. Sci., North Dakota State Univ., Tech. Rep., 2001.
[73] W. Macherey and H. Ney, "A comparative study on maximum entropy and discriminative training for acoustic modeling in automatic speech recognition," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 493–496.
[74] A. Gunawardana, M. Mahajan, A. Acero, and J. Platt, "Hidden conditional random fields for phone classification," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 1117–1120.
[75] N. Smith, M. Gales, and M. Niranjan, Data dependent kernels in SVM classification of speech patterns Univ. of Cambridge, Tech. Rep. CUED/F-INFENG/TR.387, 2001.
[76] N. Smith and M. Gales, "Speech recognition using SVMs," in *Proc. NIPS*, 2002, vol. 14, pp. 1197–1204.
[77] T. S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Proc. NIPS*, 1998, vol. 11, pp. 487–493.
[78] M. Layton and M. Gales, "Augmented statistical models for speech recognition," in *Proc. IEEE ICASSP*, France, May 2006, vol. 1, pp. 129–132.

Yasser Hifny received the M.Sc. and B.Sc. degrees in electronics and communication engineering from the University of Cairo, Cairo, Egypt, in 1995 and 2000, respectively, and the Ph.D. degree in computer science from the University of Sheffield, Sheffield, U.K, in 2006.

He is a Postdoctoral Fellow in the Human Language Technologies Group, IBM T. J. Watson Research Center, Yorktown Heights, NY. In 2000, he joined the Research and Development International (RDI), where he worked on research in text-to-speech (ArabTalk), limited domain speech compression (text-concept-to-speech), and speech verification (HAFS) projects. His research interests include speech and signal processing, machine learning, and language engineering.

Steve Renals (M'91) received the Ph.D. degree from the University of Edinburgh, Edinburgh, U.K., in 1990.

He is a Professor of Speech Technology in the School of Informatics and director of the Centre for Speech Technology Research, University of Edinburgh. He has held postdoctoral fellowships at the International Computer Science Institute, Berkeley, CA (1991–1992) and at the University of Cambridge, Cambridge, U.K. (1992–1994), and was a Member of Academic Staff at the University of Sheffield for nine years from 1994 to 2003. His research interests are in spoken language processing and multimodal interaction. He has about 150 publications in these areas.

He is an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS and of the *ACM Transactions on Speech and Language Processing*, and was previously a member of the IEEE SPS Technical Committee on Machine Learning and Signal Processing.

Authorized licensed use limited to: The University of Edinburgh. Downloaded on February 10, 2009 at 12:51 from IEEE Xplore. Restrictions apply.