Recent Developments in Large Vocabulary Continuous Speech Recognition

George Saon* and Jen-Tzung Chien†
* IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598
† Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan
gsaon@us.ibm.com & jtchien@nctu.edu.tw

***Abstract*— This paper overviews a series of recent approaches to front-end processing, acoustic modeling, language modeling, and back-end search and system combination which have made contributions for large vocabulary continuous speech recognition (LVCSR) systems. These approaches include the feature transformations, speaker-adaptive features, and discriminative features in front-end processing, the feature-space and model-space discriminative training, deep neural networks, and speaker adaptation in acoustic modeling, the backoff smoothing, large-span modeling, and model regularization in language modeling, and the system combination, cross-adaptation, and boosting in search and system combination. Some future directions for LVCSR research are also addressed.**

I. INTRODUCTION

Over the past decade, several advances have been made to the design of modern LVCSR systems to the point where their application has broadened from early speaker-dependent dictation systems to speaker-independent automatic broadcast news transcription and indexing, lectures and meetings transcription, conversational telephone speech transcription, open-domain voice search, medical and legal speech recognition and call center applications to name a few. The commercial success of these systems is an impressive testimony to how far research in LVCSR has come and the aim of this paper is to describe some of the technological underpinnings of modern systems. It must be said however that, despite the commercial success and widespread adoption, the problem of large vocabulary speech recognition is far from being solved: background noise, channel distortions, foreign accents, casual and disfluent speech or unexpected topic change can cause automated systems to make egregious recognition errors. This is because current LVCSR systems are not robust to mismatched training and test conditions and cannot handle context as well as human listeners despite being trained on thousands of hours of speech and billions of words of text.

Technological improvements have been made in four components of an LVCSR system: *front-end processing*, *acoustic modeling*, *language modeling*, *hypothesis search and system combination*. A comprehensive survey of early LVCSR systems was presented in [35]. The state of the art in LVCSR has shifted considerably since then through the advent of powerful speaker adaptation, discriminative training and language modeling techniques. This paper reports some advanced developments which are a substantial step toward making a number of high-utility applications possible [28].

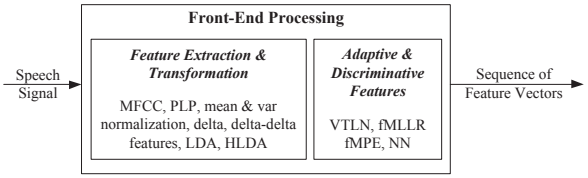

Fig. 1. Overview of front-end processing methods.

II. FRONT-END PROCESSING

A. Feature Extraction and Transformation

We first address some new front-end processing methods for LVCSR as summarized in Figure 1. The role of the front-end processing module is to extract a sequence of acoustic feature vectors from the speech waveform. Nowadays, this is done by computing a short-term fast Fourier transform (FFT) of the speech signal within a 25 msec time window 100 times per second. The energies of the neighboring frequencies within each frame are binned together via a mel-scale filterbank. Next, the log mel-spectra are decorrelated via a discrete cosine transform resulting in a 13-dimensional vector of mel frequency cepstral coefficients (MFCC). Lately, MFCCs have been replaced with a more noise-robust representation based on perceptual linear prediction (PLP) coefficients [14].

Feature extraction has benefited from the advent of two important techniques. The first is the use of utterance-based cepstral mean subtraction (CMS) and speaker-based cepstral variance normalization (CVN). The second idea has to do with incorporating temporal context across cepstral frames based on the delta and delta-delta coefficients [9]. This method has been replaced in modern LVCSR systems by a linear projection matrix which maps the vector obtained by concatenating consecutive frames to a lower-dimensional space. The projection is designed such as to maximally separate the phonetic classes in the transformed space and remove the equal class covariance constraint such as heteroscedastic linear discriminant analysis (HLDA) [17]. The LDA feature space is "rotated" by means of a semi-tied covariance transform (STC) [10] which aims to minimize the loss in likelihood between full and diagonal covariance Gaussians.

B. Speaker-Adaptive Features

The variation of the acoustic features has two components: an *intra-speaker* component due to the different phonetic classes being uttered and an *inter-speaker* component due

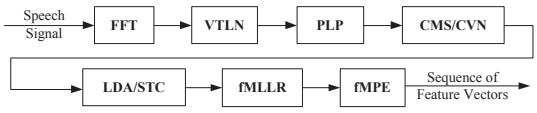

Fig. 2. Overview of front-end pipeline processing steps.

to the different vocal characteristics of the various speakers. Speaker normalization techniques operating in the feature domain aim at producing a canonical feature space by eliminating the inter-speaker variability. Examples of such techniques are: (i) warping the frequency axis to match the vocal tract length of a reference speaker as in vocal tract length normalization (VTLN) [34], and (ii) affinely transforming the features to maximize the likelihood under the current model as in feature-space maximum likelihood linear regression (fMLLR) [10].

C. Discriminative Features

Another powerful tool in the modeling arsenal of modern LVCSR systems is feature-space discriminative training. Feature-space minimum phone error (fMPE) [21] is a transformation that provides time-dependent offsets to the regular feature vectors. The offsets are obtained by a linear projection from a high-dimensional space of Gaussian posteriors. The projection is trained such as to enhance the discrimination between correct and incorrect word sequences. Another promising tack for discriminative feature extraction is the use of a neural network (NN) parameterization of the speech signal. The approach consists in estimating phone posteriors using a multi-layer perceptron and in modeling the outputs of the network with Gaussian mixture models (GMMs). A refinement to this technique was presented in [13] where bottleneck features are introduced for improving LVCSR and are derived from a 5-layer NN with a constriction in the middle (hidden layer with few units). Figure 2 illustrates the typical front-end pipeline of a modern LVCSR system.

III. ACOUSTIC MODELING

Hidden Markov models (HMMs) are a popular formalism for representation of temporal or spatial sequence data. Assume that a set of D-dimensional continuous-valued speech feature vectors $X = \{\mathbf{x}_t\}_{t=1}^{T}$ is collected for estimation of HMM parameters $\Lambda = \{\omega_{ik}, \mu_{ik}, \Sigma_{ik}\}$ consisting of mixture weights ω_{ik}, mean vectors μ_{ik} and covariance matrices Σ_{ik} for state i and GMM component k. Conventional HMMs are generative models trained according to the maximum likelihood (ML) criterion through maximizing the joint likelihood function $p(X|\Lambda)$. Figure 3 displays an overview of state-of-the-art acoustic modeling techniques for LVCSR including discriminative training and speaker adaptation.

A. Discriminative Training

ML estimation guarantees the "optimality" in distribution for a generative model. However, for LVCSR, the "optimality" in classification accuracy is desired. Discriminative estimation is more effective than ML estimation. We aim to find the best discriminative acoustic model to achieve the lowest word error

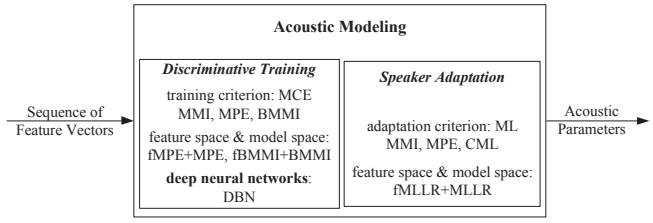

Fig. 3. Overview of acoustic modeling techniques.

rates (WERs) on unseen test data. A useful solution is to estimate the discriminative model by minimizing the classification error rate (MCE) which is a smooth approximation to the word or sentence error rate [16]. Alternatively, discriminative acoustic models can be trained according to the maximum mutual information (MMI) criterion which is expressed as the mutual information between the observation data X and the sequence of reference words W^r or equivalently as the difference between a numerator function $F^{\text{num}}(\Lambda)$ corresponding to the reference word sequence W^r and a denominator function $F^{\text{den}}(\Lambda)$ for all possible word sequences $\{W\}$ which is approximated by the sum to only the word sequences that occur in a word lattice of alternative sentence hypotheses. The MMI estimation of HMM parameters Λ is typically performed through an extended Baum-Welch algorithm by maximizing the "weak-sense" auxiliary function where an additional smoothing function is introduced to guarantee that the auxiliary function increases after parameter updates [20]. MMI training can be interpreted as a maximization of the log posterior probability $\log p_\Lambda(W^r|X)$ of the correct word sequence W^r [20] which is also known as conditional maximum likelihood (CML) estimation.

In another approach, discriminative training based on the criterion of minimum phone error (MPE) [20] aims to minimize the weighted phone error rate or equivalently maximize the weighted phone accuracy where the weighting function $A(W, W^r)$ is determined by the number of correct phones in W (given reference word sequence W^r). In addition to model-space discriminative training, the same objective function, either MPE or MMI, can be optimized to perform *feature-space* discriminative training [21]. More concretely, feature-space MPE (fMPE) or feature-space MMI (fMMI) training is performed by transforming acoustic features $\mathbf{x}_t$ to $\hat{\mathbf{x}}_t = \{\hat{x}_{td}\}$ for each frame t by $\hat{\mathbf{x}}_t = \mathbf{x}_t + M\mathbf{h}_t$ where $M = \{m_{dj}\}$ is a transformation matrix and $\mathbf{h}_t = \{h_{tj}\}$ is a high dimensional feature vector which is formed by Gaussian posteriors given the current frame and is calculated from a GMM. The transformation matrix M is estimated by maximizing the same criterion as in MPE or MMI by using a gradient descent algorithm. On several LVCSR tasks, fMPE training outperformed MPE training. The system performance was further improved by combining fMPE training with MPE training of the model parameters (also denoted by fMPE+MPE) [21].

In yet another approach inspired by large margin classification techniques, a boosted MMI (BMMI) objective function was constructed by introducing a boosting factor which is

controlled by a scaling parameter and a phone accuracy measure $A(W, W^r)$ between hypothesized and reference word sequences $\{W, W^r\}$. The underlying idea of BMMI training is to artificially increase the likelihood of more confusable sentences that have more errors so that the training algorithm focuses more on them. Feature-space and model-space BMMI training (denoted by fBMMI+BMMI) has been shown to be superior to fMPE+MPE for several LVCSR tasks [22].

Moreover, the deep neural network acoustic model was known as discriminative model and was recently popular for LVCSR with significant improvement over discriminatively-trained HMMs with state-dependent GMMs [7][30]. The moniker "deep" comes from using more than one hidden layer, typically three to five. The deep belief network (DBN) models the context-dependent output distributions directly and uses a greedy, layer-wise pretraining of the weights with either a supervised or unsupervised criterion [7]. This pretraining step prevents the supervised training of the network from being trapped in a poor local optimum.

B. Speaker Adaptation

LVCSR systems are further improved by compensating the acoustic mismatch between training and test environments via speaker adaptation by using speaker-specific data during training as well as at test time. In addition to speaker normalized feature extraction using VTLN, maximum likelihood linear regression (MLLR) [18] was developed for speaker adaptation by transforming the clusters of HMM mean vectors $\{\mu_{ik}\}$ using cluster-dependent regression matrices $M = \{M_c\}$ by $\hat{\mu}_{ik} = M_c \xi_{ik}$ where $\xi_{ik} = \left[\mu_{ik}^T \; 1\right]^T$ is an extended $(D+1)$-dimensional vector and M_c is a $D \times (D+1)$ matrix. The ML estimation of regression matrices M is formulated as closed-form solution according to an expectation-maximization (EM) algorithm [18].

Alternatively, feature space MLLR (fMLLR) [10] was proposed for speaker adaptation where the acoustic features $\{\mathbf{x}_t\}$ are transformed to $\{\hat{\mathbf{x}}_t\}$ by using a regression matrix M^f via $\hat{\mathbf{x}}_t = M^f \xi_t$ where $\xi_t = \left[\mathbf{x}_t^T \; 1\right]^T$ is an extended feature vector. The ML estimate of the regression matrix M^f is calculated by an iterative row-by-row optimization procedure. In recent LVCSR systems, acoustic models are speaker adaptively trained in a canonical feature space given by VTLN-warped and fMLLR-transformed features. At test time, speaker adaptation consists in VTLN, fMLLR and MLLR. This recipe for feature-space and model-space speaker adaptation has led to significant gains in LVCSR performance.

Speaker adaptation can be upgraded by extending generative linear transformations to *discriminative linear transformations*. MMI-based discriminative adaptation [12] was proposed to estimate the regression matrix M by maximizing the mutual information $I_\Lambda(X, W^r; M)$ given adaptation data X and reference transcription W^r. This objective function was expressed as the conditional likelihood $\log p_\Lambda(W^r|X, M)$. The CML linear regression adaptation [12] obtained good performance for LVCSR. By modifying the objective function from MMI to MPE, the phone accuracy $A(W, W^r)$ of the adaptation data is incorporated into the "weak-sense" auxiliary function. MPE-based speaker adaptation outperformed MMI-based speaker adaptation on several LVCSR tasks [33].

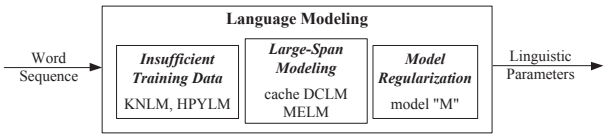

Fig. 4. Overview of language modeling methods.

IV. LANGUAGE MODELING

A statistical language model (LM) $p_\Gamma(W)$ with n-gram parameters Γ represents the prior probability of a word string $W = \{w_1, \cdots, w_T\} \triangleq w_1^T$ which is calculated by multiplying the probabilities of a predicted word w_i conditioned on the preceding $n-1$ words w_{i-n+1}^{i-1}. The prior probability $p_\Gamma(W)$ is combined with the acoustic likelihood function $p_\Lambda(X|W)$ given HMM parameters Λ to find the most likely word sequence $\hat{W}$ according to the Bayes decision rule $\hat{W} = \arg\max_W p_\Lambda(X|W)\, p_\Gamma(W)$. Although n-grams are effective at exploiting local lexical regularities, they suffer from the inadequacies of *training data*, *long-distance information* and *model generalization*, which constrain the prediction capability. Figure 4 summarizes some new language modeling methods which have been dominant for LVCSR.

A. Insufficient Training Data

Chen [4] surveyed a series of smoothing techniques of *n*-gram language model which are used to tackle the issue of inadequate training data. These techniques basically cope with zero probability estimates for *n*-grams not observed in the training corpus. Among these techniques, a variant of Kneser-Ney (KN) smoothing outperformed all other algorithms for LVCSR. The interpolated KN (IKN) smoothing was formed by utilizing absolute discounting, modified counts for *n*-gram probabilities, and interpolation with lower-order *n*-gram probabilities [4]. The discount parameter depends on the length of context w_{i-n+1}^{i-1}. A modified KN (MKN) language model [4] was proposed by extending IKN language model via allowing different discount parameters for n-grams with different counts. MKN language model outperformed IKN language model in [4].

KN language model (KNLM) was further generalized to a hierarchical Pitman-Yor language model (HPYLM) [32] where a nonparametric prior based on a Pitman-Yor (PY) process was introduced to interpret language model smoothing from a Bayesian perspective. Interpolating with lower-order n-grams is equivalent to performing hierarchical Bayesian framework by recursively combining the $(n-1)$th-order PY process priors over the nth-order predictive distributions until the unigram model is reached. PY process is a generalization of a Dirichlet process with an additional discount parameter for language model smoothing. This process produces the *power-law* distributions which are well-suited to model word frequencies in natural language [32]. Gibbs sampling can

be applied for model inference based on Chinese restaurant metaphor. HPYLM is a *Bayesian generalization* of KNLM with an additional strength parameter. In [15], HPYLM had improved performance over KNLM for LVCSR based on several large-scale training datasets.

B. Large-Span Modeling

To compensate the inadequate handling of long-distance information in n-gram models, the latent semantic analysis (LSA) was explored for construction of large-span language models. The semantic information was represented in low dimensional vector space consisting of latent topics shared for words and documents [1]. LSA language model was calculated by cosine similarity measure between a predicted word w_i and its history context w_{i-n+1}^{i-1} in the common semantic space. Integrating LSA language models with standard n-gram models has led to good LVCSR performance [1]. However, LSA cannot be generalized for unseen test data.

To tackle the generalization issue, Blei [2] presented the latent Dirichlet allocation (LDA) where Dirichlet priors were introduced to represent topic mixtures for seen and unseen documents. A history-based LDA language model was developed to calculate the n-gram probability [6]. The sequence of history words w_{i-n+1}^{i-1} is first transformed to topic space or class space via a linear function. This transformation is used to find class-dependent hyperparameters of Dirichlet priors which draw the classes for a predicted word w_i. A class mixture model is established by integrating C class distributions associated with word w_i. The resulting Dirichlet class language model (DCLM) parameters are estimated by maximizing the marginal likelihood of n-gram events over classes and class mixtures through the variational Bayes EM algorithm. DCLM was extended to a cache DCLM by combining the class information outside n-gram context w_{i-n+1}^{i}.

The maximum entropy (ME) approach is also proposed to integrate the sources of low-order n-gram, high-order n-gram, long-distance information and syntactic/semantic knowledge in an ME language model (MELM) [5]. MELM is expressed as a log linear model. Assuming that there are F features $\{f_k(\cdot)\}$ induced by the words preceding word w_i in the corresponding sentence $W^{r,i}$, ME principle is used to estimate the parameters $\{\lambda_k\}$ with maximum entropy, randomness or smoothness while all feature functions are constrained. This ME technique acts as a model smoothing method over different backoff models.

C. Model Regularization

ME model is known as an exponential n-gram model. Chen [3] addressed the issue of model regularization and investigated a variety of exponential language models to find an empirical relationship between training set cross-entropy H_{train} and test set cross-entropy H_{test} as $H_{\text{test}} \approx H_{\text{train}} + (\gamma/N_n)\sum_{k=1}^{F}|\tilde{\lambda}_k|$ where N_n is the number of n-gram events, $\tilde{\lambda} = \{\tilde{\lambda}_k\}$ are regularized ME parameters and γ is a constant independent of data and model. This relationship was used to motivate a *heuristic* for improving LVCSR performance of test data by penalizing large-sized language model with large $\tilde{\lambda}_k$ values. The heuristic was to identify groups of features with similar $\tilde{\lambda}_k$ values and add new features that were the sums of the original features in individual groups. The size of the exponential language model $\sum_{k=1}^{F}|\tilde{\lambda}_k|$ was surprisingly reduced and the prediction performance was improved [3]. The heuristic was further applied to shrink exponential language model and build a middle-sized class-based language model, called model "M", which was both smaller than the baseline classed-based model and had a lower training set cross-entropy. Model generalization was improved. Model "M" has been successfully applied in IBM systems that were fielded in LVCSR evaluations with good performance [3].

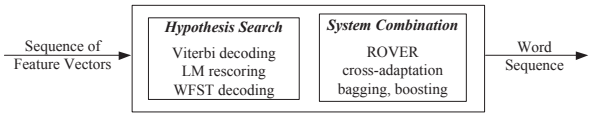

Fig. 5. Overview of hypothesis search and system combination methods.

V. HYPOTHESIS SEARCH AND SYSTEM COMBINATION

A. Hypothesis Search

Figure 5 shows several new methods which have significant impact on LVCSR decoding and system combination. A survey of early LVCSR decoders can be found in [35]. Since then, advances in decoding algorithms coupled with the availability of increased computing power has made accurate, real-time LVCSR possible. Chief among these advances is the use of weighted finite-state transducers (WFSTs) which allow to efficiently encode all the various knowledge sources present in a speech recognition system (language model, pronunciation dictionary, context decision trees and HMM topologies). The network resulting from the composition of these WFSTs, after minimization, can be directly used in a time-synchronous Viterbi decoder [19]. One such example of a WFST decoder [25] operates on static graphs obtained by successively expanding the words in an n-gram model in terms of their pronunciation variants, the phonetic sequences of these variants and the context dependent acoustic realizations of the phones. This can be done even for large cross-word phonetic contexts such as pentaphones (or quinphones). In order to use the full language model, a static decoder needs to first generate lattices with a smaller LM then rescore them with the full LM, which requires additional computations during the search.

B. System Combination

Modern LVCSR systems employ multiple decoding and rescoring passes with several speaker adaptation passes in-between. System performance can be improved through *cross-adaptation* where the output of one system is used to adapt the acoustic models of another system. Another form of system combination pioneered by ROVER [8] consists in aligning the word hypotheses from the different systems and in outputting the words which have the most votes within each bin. In

yet another approach, the lattices from multiple systems are intersected using WFST operations [19].

The acoustic models which are combined usually differ in one or more design parameters such as input features, acoustic modeling paradigm, phonetic context, discriminative training criterion. A lot of human intervention is required in choosing which systems are good for combination. Ideally, one would want an automatic procedure for training accurate systems or models which make complementary recognition errors. One such approach is a classifier combination technique called *bagging* and consists in training an ensemble of acoustic models by randomizing the questions in the context decision trees [31]. Another approach is to iteratively train a sequence of acoustic models on re-weighted training samples where the weights of incorrectly decoded frames is progressively increased. This is an adaptation of the classifier combination technique called *boosting* and has been shown to be superior to bagging for LVCSR [29]. In what follows, we present some new methods and point out possible directions for LVCSR.

VI. SOME NEW DIRECTIONS

A. Structural State Models

In general, speech feature vectors $\mathbf{x}_t$ are modeled by context dependent GMMs conditioned on HMM states and are assumed to be conditionally independent from one another. Each state has its own model parameters and there is no sharing across states. Povey [23] presented the subspace Gaussian mixture models (SGMMs) to allow all phonetic states to share a common GMM structure but with means and mixture weights varying in a subspace of the entire parameter space. The state observation distribution of feature vector $\mathbf{x}_t$ at state i is expressed by a mixture of sub-state distributions each with a mixture of GMMs

$$p_{\text{SGMM}}\left(\mathbf{x}_t|\Lambda_i\right) = \sum_{j=1}^{N_i} c_{ij} \left[\sum_{k=1}^{K} \omega_{ijk} \mathcal{N}(\mathbf{x}_t; \mu_{ijk}, \Sigma_k) \right]. \quad (1)$$

Each GMM consists of state and sub-state dependent mixture weights $\omega_{ijk} = \exp\left(\mathbf{w}_k^T \mathbf{v}_{ij}\right) / \sum_{k'=1}^{K} \exp\left(\mathbf{w}_{k'}^T \mathbf{v}_{ij}\right)$, mean vectors $\mu_{ijk} = \Phi_k \mathbf{v}_{ij}$ and canonical covariance matrices Σ_k. There are K canonical states with parameters $\{\Phi_k, \mathbf{w}_k, \Sigma_k\}$ and N_i sub-states for state i with each sub-state having its own mixture weight c_{ij} and subspace vector $\mathbf{v}_{ij}$. SGMM parameters $\Lambda_{\text{SGMM}} = \{\Lambda_{ij} = \{c_{ij}, \mathbf{v}_{ij}\}, \Lambda_k = \{\Phi_k, \mathbf{w}_k, \Sigma_k\}\}$ are estimated according to ML criterion. Compared to HMMs, a much more compact representation is obtained by SGMMs due to the canonical parameters Λ_k globally shared across the different states i and sub-states j.

SGMMs were further generalized to canonical state models (CSMs) [11] where the context-dependent transform parameters Λ_{ij} and the canonical state model parameters Λ_k are involved in the state likelihood calculation. The context-dependent state parameters are a transformed version of one or more canonical state parameters which represents the sub-state parameters of a Markov state. The state likelihood of $\mathbf{x}_t$ given a context-dependent state i is similar to (1) except that the mixture weights, mean vectors and covariance matrices of the GMM are replaced by general transformation functions $\omega_{ijk} = F_\omega\left(k, \theta_{ij}\right)$, $\mu_{ijk} = F_\mu\left(k, \theta_{ij}\right)$ and $\Sigma_{ijk} = F_\Sigma\left(k, \theta_{ij}\right)$, respectively, where θ_{ij} denotes the set of transform parameters, c_{ij} is seen as the transform prior and $\Lambda_{ij} = \{c_{ij}, \theta_{ij}\}$. This CSM is a general model and can be realized to the mixtures of MLLR transforms, mixtures of fMLLR transforms and SGMMs which differ in $F_\omega(\cdot)$, $F_\mu(\cdot)$ and $F_\Sigma(\cdot)$ that are applied to map the canonical state k to the context-dependent state i [11]. SGMMs and CSMs have been successfully applied to several LVCSR tasks [23][11].

B. Basis Representation

LVCSR systems are usually constructed by collecting large amounts of training data and estimating a large number of model parameters to achieve desirable recognition accuracy on test data. A large set of context-dependent Gaussian components is trained. However, GMMs may not be an accurate representation of high dimensional acoustic features. Alternatively, acoustic feature vectors can be viewed as lying in a vector space spanned by a set of basis vectors. Such a basis representation has been popular in the fields of machine learning and signal processing. This direction is now increasingly important for acoustic feature representation [24].

Bayesian sensing HMMs (BS-HMMs) [27] were developed by incorporating Markov chains into the basis representation of continuous speech. The underlying aspect of BS-HMMs is to measure an observed feature vector $\mathbf{x}_t$ based on a compact set of state-dependent dictionary $\Phi_i = [\varphi_{i1}, \cdots, \varphi_{iN}]$. The reconstruction error between measurement $\mathbf{x}_t$ and its representation $\Phi_i \mathbf{w}_t$, where $\mathbf{w}_t = [w_{t1}, \cdots, w_{tN}]^T$, is assumed to be Gaussian distributed with zero mean and a state-dependent precision matrix R_i. Bayesian sensing is to yield "distribution estimates" of the speech feature vectors due to the variations of sensing weights $\mathbf{w}_t$. A Gaussian prior with zero mean and state-dependent diagonal covariance matrix is introduced to characterize the weight vector, i.e. $\mathcal{N}\left(\mathbf{w}_t; 0, \text{diag}\left\{\alpha_{in}^{-1}\right\}\right)$. This prior is prone to be sparse [27]. The automatic relevance determination (ARD) parameters $\{\alpha_{in}\}$ are likely to be large to draw zero values for $\mathbf{w}_t$. Only relevant basis vectors are selected to represent sequence data. BS-HMM parameters are formed by $\Lambda_{\text{BSHMM}} = \{\Phi_i, A_i, R_i\}$ where their implicit solutions were derived by EM algorithm according to the ML type II criterion [27]. The state likelihood, marginalized over sensing weights, was illustrated as a new Gaussian distribution with a factor analyzed covariance matrix [26]. In the latest DARPA GALE Arabic broadcast news transcription evaluation, BS-HMMs trained on 1800 hours of data outperformed state-of-the-art HMMs even after feature-space and model-space discriminative training [26].

C. Model Regularization

ML acoustic models and language models in LVCSR systems may suffer from an over-training problem where the estimated models are too complex to generalize for future data. This leads to a limited prediction capability on unknown

test sentences. Also, the real-world continuous speech is collected from heterogeneous environments with mismatched training and test conditions and various variations due to noise, channel, gender, speaker, accent, co-articulation, speaking rate, emotion, etc. The issues of overtraining and heterogeneous data warrant more investigation. In addition, training data may be incorrectly labeled or even without labels. The selected model structure may not be appropriate for the collected data or the assumed models may be different from the true ones. Estimation errors may exist in the model construction due to sparse data, approximate inference or slow convergence. Overall, future LVCSR should tackle model regularization and compensate for the *uncertainties* in the construction of component models. Model-space and feature-space speaker adaptation provides a solution to regularize the trained model for test conditions. The language model based on model "M" [3] and the acoustic model based on BS-HMMs [27] are two new trends towards high-performance LVCSR as far as model regularization is concerned. Nevertheless, there are other LVCSR components which have not been thoroughly investigated from the perspective of model regularization.

VII. CONCLUSIONS

We have surveyed a series of approaches to front-end processing, acoustic modeling, language modeling and back-end search and system combination which have made big contributions for LVCSR in the past decade or so. We presented flexible acoustic models based on structural state models and robust basis representation. With the aim of modeling unknown variations in the data and model parameters, we pointed out possible future directions towards structural learning and model regularization for the different components of an LVCSR system.

REFERENCES

[1] J. Bellegarda, "Exploiting latent semantic information in statistical language modeling", *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1279–1296, 2000.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation", *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] S. F. Chen, "Shrinking exponential language models", *Proc. of NAACL-HLT*, pp. 468–476, 2009.

[4] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling", *Computer Speech and Language*, vol. 13, no. 4, pp. 359–394, 1999.

[5] J.-T. Chien, "Association pattern language modeling", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1719–1728, 2006.

[6] J.-T. Chien and C.-H. Chueh, "Dirichlet class language models for speech recognition", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 482–495, 2011.

[7] G. E. Dahl, D. Yu, L. Deng and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 30-42, 2012.

[8] J. Fiscus, "A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER)", *Proc. of ASRU*, pp. 347–354, 1997.

[9] S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, no. 34, pp. 52–59, 1986.

[10] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition", *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[11] M. J. F. Gales and K. Yu, "Canonical state models for automatic speech recognition", *Proc. of INTERSPEECH*, pp. 58–61, 2010.

[12] A. Gunawardana and W. Byrne, "Discriminative speaker adaptation with conditional maximum likelihood linear regression", *Proc. of EUROSPEECH*, pp. 1203–1206, 2001.

[13] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings", in *Proc. of ICASSP*, pp. 757–760, 2007.

[14] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *Journal of the Acoustic Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[15] S. Huang and S. Renals, "Hierarchical Bayesian language models for conversational speech recognition", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 8, pp. 1941–1954, 2010.

[16] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error methods for speech recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 257–265, 1997.

[17] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition", *Speech Communication*, vol. 26, pp. 283–297, 1998.

[18] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[19] M. Mohri, F. Perreira and M. Riley, "Weighted finite state transducers in speech recognition", *Computer Speech and Language*, vol. 16, pp. 69–88, 2002.

[20] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training", *Proc. of ICASSP*, pp. 105–108, 2002.

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

[22] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training", *Proc. of ICASSP*, pp. 4057–4060, 2008.

[23] D. Povey et al., "The subspace Gaussian mixture models - a structured model for speech recognition", *Computer Speech and Language*, vol. 25, pp. 404–439, 2011.

[24] T. N. Sainath, A. Carmi,D. Kanevsky and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification", *Proc. of ICASSP*, pp. 4370–4373, 2010.

[25] G. Saon, D. Povey and G. Zweig, "Anatomy of an extremely fast LVCSR decoder", *Proc. of INTERSPEECH*, pp. 549–552, 2005.

[26] G. Saon and J.-T. Chien, "Some properties of Bayesian sensing hidden Markov models", *Proc. of ASRU*, pp. 65-70, 2011.

[27] G. Saon and J.-T. Chien, "Bayesian sensing hidden Markov models", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 43–54, 2012.

[28] G. Saon and J.-T. Chien, "Some recent advances in large vocabulary continuous speech recognition", *IEEE Signal Processing Magazine*, November 2012.

[29] G. Saon and H. Soltau, "Boosting systems for large vocabulary continuous speech recognition", *Speech Communication*, vol. 54, no. 2, pp. 212–128, 2012.

[30] F. Seide, G. Li, X. Chen and D. Yu, "Conversational speech transcription using context-dependent deep neural networks", *Proc. of INTERSPEECH*, pp. 437-440, 2011.

[31] O. Siohan, B. Ramabhadran and B. Kingsbury, "Constructing ensembles of ASR systems using randomized decision trees" *Proc. of ICASSP*, pp. 197–200, 2005.

[32] Y. W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes", *Proc. of Annal Meeting of ACL*, pp. 985–992, 2006.

[33] L. Wang and P. C. Woodland, "MPE-based discriminative linear transforms for speaker adaptation", *Computer Speech and Language*, no. 22, pp. 256–2727, 2008.

[34] S. Wegmann, D. McAllaster, J. Orloff and B. Peskin, "Speaker normalization on conversational telephone speech", *Proc. of ICASSP*, pp. 339–341, 1996.

[35] S. Young, "A review of large-vocabulary continuous-speech recognition", *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 45–57, 1996.