George Saon and Jen-Tzung Chien

Large-Vocabulary Continuous Speech Recognition Systems

A look at some recent advances

Over the past decade or so, several advances have been made to the design of modern large-vocabulary continuous speech recognition (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 article 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.

Digital Object Identifier 10.1109/MSP.2012.2197156
Date of publication: 15 October 2012

1053-5888/12/$31.00©2012IEEE

INTRODUCTION

There is a vast body of literature on LVCSR research and some limitation is necessary in the scope of this article. We will focus primarily on the techniques that have been successful in various U.S. government-led speech recognition evaluations that aim to measure yearly progress in the field of automatic speech recognition. These techniques have been incorporated in most competition-grade LVCSR systems fielded by universities such as Cambridge (United Kingdom), LIMSI (France), RWTH Aachen (Germany), Carnegie Mellon University (United States), and commercial institutions like AT&T, BBN, IBM, and SRI in the areas of English conversational telephone speech transcription [1]–[4] and broadcast news transcription for English [3], [5], Arabic [6]–[10], and Mandarin [11]–[13]. Moreover, we will limit the discussion to language-independent techniques and will not address language specific issues such as, for example, tone modeling for Mandarin or vowelization for Arabic.

Technological improvements have been made in all areas of LVCSR: front-end processing, acoustic modeling, language modeling, hypothesis search, and system combination as shown in Figure 1. A comprehensive survey of early LVCSR systems was presented in [14] and, more recently, in [15]. 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 some of which will be detailed in this article. In [16] and [17], the grand challenges in speech recognition and understanding were addressed. Compared to [16] and [17], this article reports more advanced and focused techniques in different areas of LVCSR, which are a substantial step toward implementing these grand challenges and making a number of high-utility applications possible.

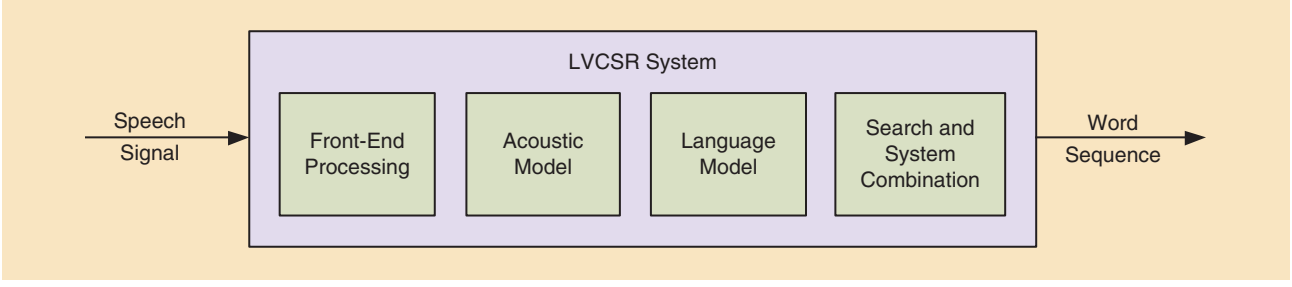
FRONT-END PROCESSING

We first address some new front-end processing methods for LVCSR that cover feature extraction and transformation, noise robust feature processing, and the estimation of adaptive and discriminative features as summarized in Figure 2.

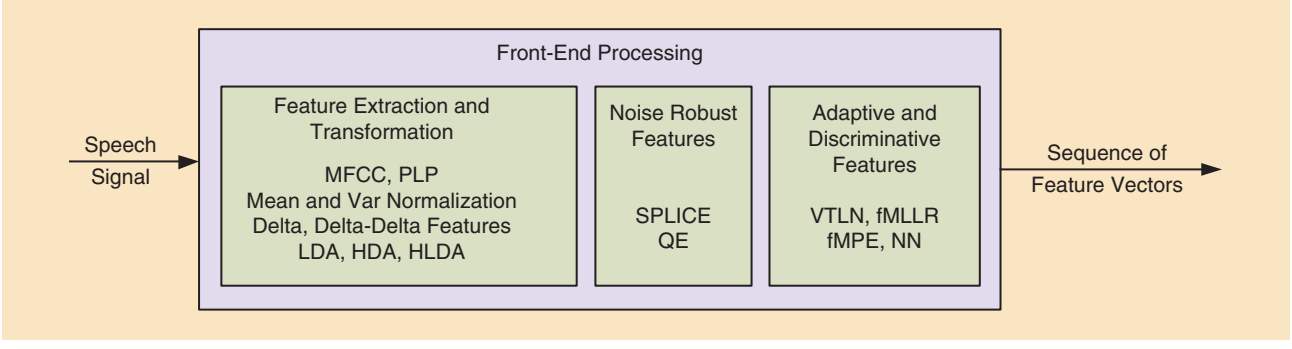
FEATURE EXTRACTION AND TRANSFORMATION

The role of the front-end processing module is to extract a sequence of acoustic feature vectors X from the speech waveform S. Today, this is done by computing a short-term fast Fourier transform (FFT) of the speech signal within a 25 ms time window 100 times/s. The energies of the neighboring frequencies within each frame are binned together via a mel-scale filterbank whose width and spacing of the filters is inspired by human auditory processing. Next, a logarithm is applied to the outputs of the filters and 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 [18].

In the context of LVCSR, feature extraction has benefited from the advent of two important techniques. The first is the use of speaker-based mean and variance normalization of the cepstral coefficients. While utterance-based cepstral mean subtraction (CMS) is a well-known technique, cepstral variance normalization (CVN) at the speaker level was introduced recently during the HUB-5 (or Switchboard) evaluations [1]. The second idea has to do with incorporating temporal context across cepstral frames. A common practice is to compute speed and acceleration coefficients (also called delta and delta-delta



[FIG1] **The four components of an LVCSR system.**



[FIG2] **Overview of front-end processing methods.**

coefficients) from the neighboring frames within a window of, typically, +/− 4 frames. These coefficients are appended to the static cepstra to form the final feature vector [19]. This ad hoc heuristic has been replaced in modern LVCSR systems by a linear projection matrix that 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. The separation is typically measured by a linear discriminant analysis (LDA) criterion [20]. Extensions of LDA, which remove the equal class covariance constraint such as heteroscedastic LDA (HLDA) [21], [22] and variants thereof [20], have also been considered. The resulting feature vectors are typically modeled with diagonal covariance Gaussians. To make the diagonal covariance modeling assumption more valid, the LDA feature space is "rotated" by means of a semitied covariance transform (STC) [23], which aims to minimize the loss in likelihood between full and diagonal covariance Gaussians. Using this cascade of LDA and STC transforms leads to a 10–15% relative improvement in word error rate (WER) over simple temporal derivatives on several LVCSR tasks [20].

NOISE ROBUST FEATURES

Speech signals are often contaminated with environmental noises, which can adversely affect the recognition performance. Developing noise robustness techniques in the front-end processing is crucial to ensure robustness in speech recognition [24]. One such algorithm, called SPLICE [25], [26], which stands for "stereo-based piecewise linear compensation for environments," was proposed for noisy speech recognition in nonstationary noise environments. The essence of the SPLICE algorithm is to perform feature enhancement by removing noise from the corrupted speech via the most likely correction vector that is the expected difference between the clean speech and the corrupted speech, associated with the most probable region in acoustic space. Stereo clean/noisy speech data are required to estimate maximum likelihood correction vectors. In [27], another algorithm called quantile-based histogram equalization (QE) was developed to compensate the mismatched distributions of training and test speech data based on the quantiles of the distributions. The parameters of a compensation function were estimated by minimizing the squared distance between the current quantiles and the training quantiles in the mel-scaled filter bank. SPLICE and QE were evaluated for noisy speech recognition using *The Wall Street Journal* (*WSJ*) corpus under various noise types and noise levels. Significant improvements were obtained in clean and multicondition training scenarios.

SPEAKER-ADAPTIVE FEATURES

The training data for a speaker independent system usually comprises speech from a large number of different speakers. The variation of the acoustic features can be seen as having two components: an intraspeaker component due to the different phonetic classes being uttered and an interspeaker component due to the different vocal characteristics of the various speakers. For the purpose of discriminating between phonetic classes, we are only interested in modeling the intraspeaker variation rather than the interspeaker variation. Speaker normalization techniques operating in the feature domain aim at producing a canonical feature space by eliminating as much of the interspeaker variability as possible. Examples of such techniques are as follows:

1) warping the frequency axis to match the vocal tract length of a reference speaker as in vocal tract length normalization (VTLN) [28], [29]
2) affinely transforming the features to maximize the likelihood under the current model as in feature-space maximum likelihood linear regression (fMLLR) [23]
3) a dimension-wise nonlinear transformation of the empirical distribution of the adaptation data to match a reference normal distribution as in feature space Gaussianization [30].

Next, the acoustic model is trained in this canonical feature space, which ideally becomes devoid of interspeaker variations. Speaker-adaptive features in combination with model-space adaptation results in performance improvements ranging from 20% to 30% relative on a variety of LVCSR tasks [1], [31].

DISCRIMINATIVE FEATURES

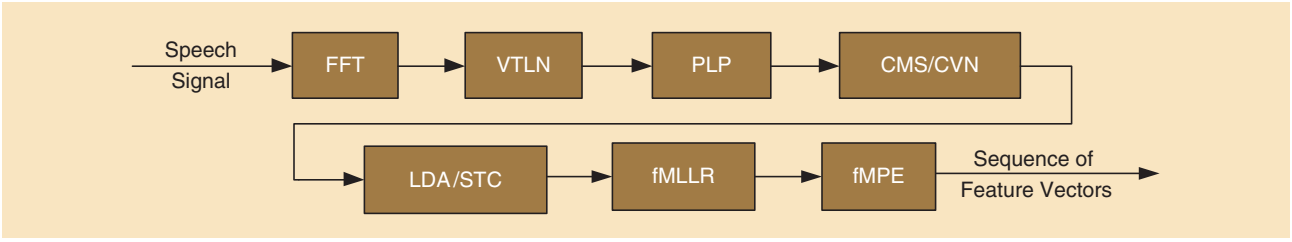
Another powerful tool in the modeling arsenal of modern LVCSR systems is feature-space discriminative training. Feature-space minimum phone error (fMPE) [32] is a transformation that provides time-dependent offsets to the regular feature vectors. The offsets are obtained by a linear projection from a high-dimensional space of Gaussian posteriors. The projection is trained such as to enhance the discrimination between correct and incorrect word sequences. In conjunction with model-space discriminative training, this technique usually leads to a 25% relative improvement in recognition performance on several tasks [32], [33]. Another promising tack for discriminative feature extraction is the use of a neural network (NN) (or connectionist) parameterization of the speech signal. The approach proposed in [34] consists in estimating phone posteriors using a multilayer perceptron and in modeling the outputs of the network with conventional Gaussian mixture models. A refinement to this technique was presented in [35] where bottleneck features are introduced for LVCSR and are derived from a five-layer NN with a constriction in the middle (hidden layer with few units). While not necessarily better by themselves, models built on NN acoustic features improve LVCSR performance through system combination [10], [36].

In summary, the typical front-end pipeline of a modern LVCSR system is illustrated in Figure 3. Next, we briefly review some fundamental acoustic modeling techniques and present some new methods that can be encountered in modern LVCSR systems.

ACOUSTIC MODELING

HIDDEN MARKOV MODELS

Hidden Markov models (HMMs) [37] are a popular formalism for the representation of temporal or spatial sequence data, e.g., speech, image, video, text, music, biology, finance, and many



[FIG3] **Overview of front-end processing steps.**

others. Assume that a set of D-dimensional continuous-valued speech feature vectors $X = \{\mathrm{x}_t\}_{t=1}^{T}$ is collected for acoustic modeling. The state observation probability density function of a feature vector x_t at time t is expressed by Gaussian mixture model (GMM)

$$p(\mathrm{x}_t|\Lambda_i) = \sum_{k=1}^{K} \omega_{ik} N(\mathrm{x}_t; \mu_{ik}, \Sigma_{ik}), \tag{1}$$

where the state parameters $\Lambda = \{\Lambda_i\} = \{\omega_{ik}, \mu_{ik}, \Sigma_{ik}\}$ consist of mixture weights ω_{ik}, mean vectors μ_{ik}, and covariance matrices Σ_{ik} for K Gaussian mixture components. Typically, Σ_{ik} are assumed to be diagonal although more sophisticated models such as subspace precision and mean (SPAM) [38] aiming to bridge the gap between full and diagonal covariance modeling have been proposed.

The joint likelihood of speech data collection X is given by

$$p(X|\Lambda) = \sum_{S=\{s_t\}} \left[\pi_{s_1} p(\mathrm{x}_1|\Lambda_{s_1}) \prod_{t=2}^{T} a_{s_{t-1}s_t} p(\mathrm{x}_t|\Lambda_{s_t}) \right]. \tag{2}$$

The HMM parameters $\Lambda = \{\pi_i, a_{ij}, \omega_{ik}, \mu_{ik}, \Sigma_{ik}\}$ obey the constraints of initial state probabilities $\sum_i \pi_i = 1$, state transition probabilities $\sum_j a_{ij} = 1$, and mixture weights $\sum_k \omega_{ik} = 1$.

MAXIMUM LIKELIHOOD ESTIMATION

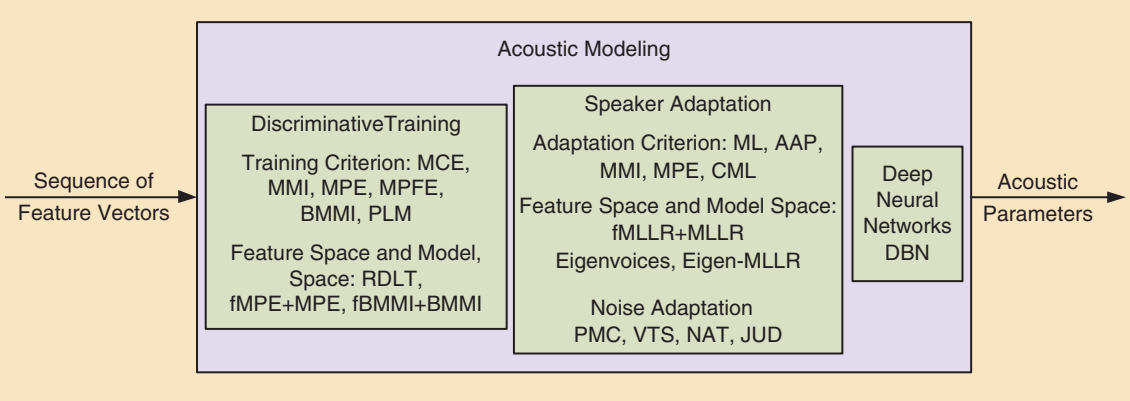
Conventional HMMs are generative models trained according to the maximum likelihood (ML) criterion where the model parameters are estimated by maximizing the joint likelihood function $p(X|\Lambda)$. ML estimation suffers from an incomplete data problem because the state labels $s_t = i$ are missing in the objective function $p(X|\Lambda)$. The expectation-maximization (EM) algorithm [39] is used to tackle this problem by maximizing the expectation function or auxiliary function of the log likelihood $\log p(X|\Lambda)$ over the missing variables $\{S = \{s_t\}\}$. At each EM iteration, a new ML estimate Λ is obtained by maximizing the auxiliary function $Q(\Lambda|\Lambda^{(k)})$ given the old estimate $\Lambda^{(k)}$ at the kth iteration

$$\begin{aligned} \Lambda^{(k+1)} &= \arg\max_{\Lambda} Q(\Lambda|\Lambda^{(k)}) \\ &= \arg\max_{\Lambda} \sum_{S} p(S|X, \Lambda^{(k)}) \log p(X, S|\Lambda). \end{aligned} \tag{3}$$

Performing EM iterations guarantees that the likelihood function does not decrease, i.e., new estimate Λ and old estimate $\Lambda^{(k)}$ satisfy $p(X|\Lambda) \geq p(X|\Lambda^{(k)})$ if $Q(\Lambda|\Lambda^{(k)}) \geq Q(\Lambda^{(k)}|\Lambda^{(k)})$ [39]. Figure 4 displays an overview of state-of-the-art acoustic modeling techniques for LVCSR. The various approaches for discriminative training, speaker adaptation, and noise adaptation are summarized by their objective functions. Several joint algorithms in feature space and model space are indicated. Acoustic modeling using deep neural networks is addressed.

DISCRIMINATIVE TRAINING

ML estimation guarantees the "optimality" in distribution for a generative model. However, for general pattern recognition systems the "optimality" in classification accuracy is desired. By being directly related to classification accuracy, discriminative estimation is more effective than ML estimation. In LVCSR systems, we aim to find the best discriminative acoustic model to achieve the lowest WERs on unseen test data. Directly



[FIG4] **Overview of acoustic modeling techniques.**

minimizing WER is hard because the objective function is not differentiable and gradient-based techniques cannot be applied. An alternative 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. MCE estimation originated from the Bayes' decision rule and significantly outperformed ML estimation for speech recognition [40]. Alternatively, discriminative acoustic models can be trained according to the maximum mutual information (MMI) criterion [41], which is expressed as the mutual information between the observation data X and the sequence of reference words W^r

$$\begin{aligned} F_{\mathrm{MMI}}(\Lambda) &\triangleq I_\Lambda(X,W^r) = \log\frac{p_\Lambda(X,W^r)}{p_\Lambda(X)p(W^r)} \\ &= \log p_\Lambda(X|W^r) - \log\sum_W p_\Lambda(X|W)\,p(W) \\ &\triangleq F^{\mathrm{num}}(\Lambda) - F^{\mathrm{den}}(\Lambda), \end{aligned} \tag{4}$$

or equivalently as the difference between a numerator function $F^{\mathrm{num}}(\Lambda)$ corresponding to the reference word sequence W^r and a denominator function $F^{\mathrm{den}}(\Lambda)$ for all possible word sequences $\{W\}$. When the exact reference is not available, the decoded output (unsupervised training) or the agreement between the decoded output and some available transcript (lightly supervised training [42]) can be substituted. The denominator term $F^{\mathrm{den}}(\Lambda)$ can be efficiently approximated by restricting the sum to only the word sequences that occur in a word lattice of alternative sentence hypotheses obtained by decoding with a weak (typically unigram) language model. The objective in (4) is similar to the negative misclassification error rate function in MCE estimation [43]. The MMI estimation of HMM parameters Λ is typically performed through an extended Baum-Welch algorithm by maximizing the "weak-sense" auxiliary function $Q(\Lambda|\Lambda^k)$ given by [44]

$$\begin{aligned} \sum_S p(S|X,W^r,\Lambda^{(k)})\log p(X,S|\Lambda) - \sum_S\sum_W p(S,W|X,\Lambda^{(k)}) \\ \times \log p(X,S|\Lambda) + Q^{\mathrm{sm}}(\Lambda|\Lambda^{(k)}), \end{aligned} \tag{5}$$

where the first and second terms correspond to the auxiliary functions for the numerator $F^{\mathrm{num}}(\Lambda)$ and denominator $F^{\mathrm{den}}(\Lambda)$, respectively, and $Q^{\mathrm{sm}}(\Lambda|\Lambda^{(k)})$ denotes a smoothing function which is added so as to guarantee that the objective function $Q(\Lambda|\Lambda^{(k)})$ increases after parameter updates. A popular smoothing function is given by the sum of negative Kullback-Leibler divergences between the state-conditional distributions for Λ and $\Lambda^{(k)}$. From (4), 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 [44], which is also known as conditional maximum likelihood (CML) estimation.

In another approach, discriminative training based on the criterion of minimum phone error (MPE) [44] has been successfully developed for LVCSR. Unlike MCE and minimum word error objective functions, which are used for minimizing the weighted sentence error rate [40] and the weighted WER [44], respectively, MPE training aims to minimize the weighted phone error rate or equivalently maximize the weighted phone accuracy

$$F_{\mathrm{MPE}}(\Lambda) \triangleq \sum_{r=1}^{R}\sum_W p_\Lambda^\kappa(W|X_r)A(W,W^r), \tag{6}$$

where $X = \{X_r\}_{r=1}^R$ denotes R training sentences, $p_\Lambda^\kappa(W|X_r)$ is defined as a scaled posterior sentence probability of hypothesized word sequence W with a scalar κ, and $A(W,W^r)$ is the number of correct phones in W (given reference word sequence W^r). MPE training leads to improved accuracy over ML and MMI training for different LVCSR tasks [44]. MPE training can be computed in a lattice framework where a lattice or word graph is generated to efficiently encode all possible word sequences which have appreciable likelihood given the acoustic evidence [45]. A variant of MPE called minimum phone frame error (MPFE) was proposed in [46] and has the advantage that it uses a frame-based phone accuracy measure (as opposed to raw phone accuracy), which is easier to compute.

In addition to model-space discriminative training for the HMM parameters Λ, the same objective function, either MPE or MMI, can be optimized to perform feature-space discriminative training, which consists in estimating a projection matrix that maps high-dimensional posterior vectors to offset vectors, which get added to the acoustic features [32]. 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 that is formed by Gaussian posteriors given the current frame and is calculated from a GMM. The transformation matrix M is estimated by maximizing the auxiliary function $Q(\Lambda|\Lambda^{(k)})$ (without the smoothing term) under the same criterion as in (4) or (6) by using a gradient descent algorithm

$$m_{dj} \leftarrow m_{dj} + \nu_{dj}\frac{\partial Q}{\partial m_{dj}} = m_{dj} + \nu_{dj}\sum_t \frac{\partial Q}{\partial \hat{x}_{td}}\,h_{tj}, \tag{7}$$

where the parameter-specific learning rate ν_{dj} is empirically determined. Since the MPE or MMI objective function depends on the HMM parameters Λ and the transformed features $\{\hat{\mathbf{x}}_t\}$, the partial differentiation in (7) contains a direct derivative and an indirect derivative

$$\frac{\partial}{\partial \hat{\mathbf{x}}_t}Q(\hat{\mathbf{x}}_t, \boldsymbol{\lambda}(\hat{\mathbf{x}}_t)|\Lambda^{(k)}) = \underbrace{\frac{\partial Q}{\partial \hat{\mathbf{x}}_t}}_{\text{direct}} + \underbrace{\frac{\partial Q}{\partial \Lambda}\frac{\partial \Lambda}{\partial \hat{\mathbf{x}}_t}}_{\text{indirect}}, \tag{8}$$

which is detailed in [32]. Observe that fMPE can be equivalently written as a mixture of time-dependent biases $\hat{\mathbf{x}}_t = \sum_j h_{tj}(\mathbf{x}_t + \mathbf{m}_j)$ where h_{tj} is the posterior for Gaussian j at time t and $\mathbf{m}_j$ is the jth column of M. A generalization of fMPE called region-dependent linear transform (RDLT) was introduced in [47] and consists of replacing the biases with a mixture of affine transforms $\hat{\mathbf{x}}_t = \sum_j h_{tj}(A_j\mathbf{x}_t + \mathbf{b}_j)$. 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) [32].

In yet another approach inspired by large-margin classification techniques, a boosted MMI (BMMI) objective function was constructed by introducing a scaling parameter κ and a boosting factor into the MMI objective function in (4) as follows [33]:

$$F_{\text{BMMI}}(\Lambda) \triangleq \sum_{r=1}^{R} \log \frac{p_\Lambda^\kappa(X_r|W^r)p(W^r)}{\sum_W p_\Lambda^\kappa(X_r|W)p(W)\exp(-bA(W,W^r))}. \tag{9}$$

The boosting factor is controlled by parameter b and phone accuracy measure $A(W,W^r)$ between hypothesized word sequence and reference word sequence (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 [11], [31], [33] and is currently the best discriminative training scheme for LVCSR for which we are aware.

To make the link with large margin classification more explicit, in [48] and [49] the BMMI criterion was modified as

$$F_{\text{PLM}}(\Lambda,b) \triangleq b + \frac{1}{\rho}\sum_{r=1}^{R} \log \frac{p_\Lambda(X_r|W^r)p(W^r)}{\sum_W p_\Lambda(X_r|W)p(W)\exp(bH(W,W^r))}, \tag{10}$$

which is seen as a penalized large-margin (PLM) criterion and is derived from a constrained optimization problem for general large-margin classification

$$\begin{aligned} &\max\ b \\ &s.t.\quad \log p_\Lambda(X_r,W^r) - \log p_\Lambda(X_r,W) \geq bH(W,W^r), \\ &\qquad\qquad \forall W, 1 \leq r \leq R. \end{aligned} \tag{11}$$

In (10) and (11), $H(W,W^r)$ denotes the number of frame phone errors or the Hamming distance between W and W^r [50], $b \geq 0$ is viewed as a margin scale parameter, and ρ is a penalty parameter controlling the tradeoff between margin maximization and constraints. The tradeoff parameter is similar to that adopted in soft-margin classification [51], [52] where the soft margin is proportional to the number of errors in a hypothesized sentence. In [52] and [53], large-margin estimation was proposed by performing frame selection and utterance selection. Support tokens for acoustic modeling are identified similar to the support vectors used in a support vector machine. In [54], Bayesian large-margin estimation was proposed by combining Bayesian learning and large-margin estimation for HMM training and model regularization. Compared to the overview of discriminative training methods in [55], this article additionally surveys feature-space discriminative training and large-margin training based on fMPE, fMPE+MPE, BMMI, fBMMI+BMMI, and PLM, which were effective in improving LVCSR performance.

SPEAKER ADAPTATION

Speaker adaptation aims to compensate the acoustic mismatch between training and test environments and is playing an important role in LVCSR systems. System performance is improved by conducting speaker adaptation during training as well as at test time by using speaker-specific data. In [28] and [29], vocal tract length normalization (VTLN) was proposed to reduce the variability among speakers. The basic idea of VTLN is to determine speaker-specific warp scales of the frequency axis and to normalize the speech signal from all speakers to that of a single canonical speaker with a standard vocal tract length. A generic speech model is iteratively trained from voiced frames of warped data and is employed to select the updated warp scale [29]. VTLN can be applied prior to extracting PLP [18] cepstral features on a per-speaker basis.

In addition to speaker normalized feature extraction, maximum likelihood linear regression (MLLR) [56] was developed for speaker adaptation by maximizing the likelihood of the adaptation data X given the correct word sequence for supervised MLLR or given the decoded word sequence or a lattice of word sequences for unsupervised MLLR [57]. Generally, supervised adaptation is performed by using training utterances or enrollment data from the same speaker and unsupervised adaptation (or self adaptation) is done on the test utterances. MLLR is a transformation-based adaptation technique where clusters of speaker-independent HMM Gaussian mean vectors $\{\mu_{ik}\}$ are transformed using cluster-dependent regression matrices $M = \{M_c\}$ by $\hat{\mu}_{ik} = M_c\xi_{ik}$ where $\xi_{ik} = [\mu_{ik}^T\ 1]^T$ is an extended $(D+1)$-dimensional vector and M_c is a $D\times(D+1)$ matrix. Similar to the ML estimation of HMM parameters Λ in (3), the ML estimation of regression matrices M is formulated according to an EM algorithm where the auxiliary function $Q(M|M^{(k)})$ of the log likelihood $\log p_\Lambda(X|M)$ of the new estimate M given the old estimate $M^{(k)}$ at iteration k is maximized. The row vectors of $M = \{M_c\}$ that maximize $Q(M|M^{(k)})$ have a closed-form solution.

Alternatively, fMLLR [23] 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 $\xi_t = [\mathbf{x}_t^T\ 1]^T$ is an extended feature vector. The ML estimate of the regression matrix M^f is calculated according to a new auxiliary function $Q(M^f|M^{f(k)})$ based on the likelihood of the transformed adaptation data $\hat{\mathbf{x}}_t$ plus the log determinant $\log|\det(M^f)|$ due to the Jacobian of the feature-space transformation. Unlike MLLR, there is no closed-form solution for fMLLR; an iterative row-by-row optimization of M^f has to be performed as shown in [23]. Traditionally, MLLR and fMLLR are derived by assuming diagonal covariance matrices $\Sigma_{ik} = \text{diag}\{\sigma_{ik}^2\}$. Efficient solutions to MLLR and fMLLR for full covariance matrices were presented in [58]. In LVCSR systems [1], [9], [36], [59], [60], 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.

In the context of rapid speaker adaptation, one technique aptly named eigenvoices [61] assumes that the supervector of Gaussian means lies in a subspace spanned by a few eigenvectors and the adaptation consists of estimating the coefficients of the linear expansion. In a similar vein, eigen-MLLR [62] considers the adaptation matrix to be a linear combination of eigen-matrices.

Speaker adaptation can be improved by extending generative linear transformations such as MLLR to discriminative linear transformations trained using discriminative criteria like aggregate a posteriori (AAP) [63], MMI [64], and MPE [65]. The AAP criterion was established by aggregating or summing up the posterior probabilities of all the classes at the sentence level. A closed-form solution to AAP regression was derived in [63] and was shown to be faster than MCE-based linear regression [66], where the generalized probabilistic descent algorithm was applied to iteratively estimate the regression parameters. MMI-based discriminative adaptation 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 logarithm of the a posteriori probability or the conditional likelihood $\log p_\Lambda(W^r|X, M)$. The CML linear regression adaptation [64] was performed after several EM iterations of MLLR adaptation (denoted by MLLR+CMLLR). Good improvements on supervised and unsupervised adaptation were obtained 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 as shown in (5) and (6). MPE-based speaker adaptation outperformed MMI-based speaker adaptation on several LVCSR tasks [65].

NOISE ADAPTATION

Some model-based compensation methods were also proposed for noise robust speech recognition on large-vocabulary speech corpora under different added noises and signal-to-noise ratios. Parallel model combination (PMC) [67] was developed for robust continuous speech recognition where the corrupted speech due to additive noise was compensated by combining clean speech HMMs and noise HMMs in the log spectral domain via mismatch functions for static and dynamic ceptral features. Vector Taylor series-based (VTS) compensation [68] was presented to adapt the existing HMM parameters to an unknown noisy environment. This HMM adaptation technique is based on a first-order Taylor series expansion around the Gaussian means of the clean speech signal and noise signals where the additive noise and the convolutive noise are characterized by an acoustic environmental model. To handle different noise environments, the noise adaptive training (NAT) [25] algorithm was proposed as a form of multistyle training where the pools of noisy speech with different noise types and noise levels are preprocessed by SPLICE-based noise reduction and then applied to estimate an integrated set of HMMs that was robust across a wide range of noise conditions. In addition, noise robustness in speech recognition was achieved by exploiting the uncertainty introduced by noise interference. The issue of uncertainty was tackled during recognition by adding the variance of the error due to feature enhancement to the HMM variances. Joint uncertainty decoding (JUD) [26] was implemented according to a joint Gaussian density of the clean and corrupted speech.

DEEP NEURAL NETWORKS

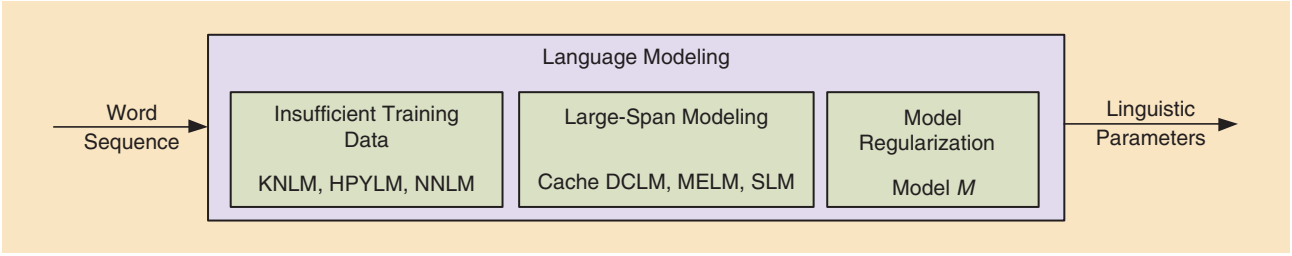
For the past 30 years or so, HMMs with state-dependent GMMs have been the de facto standard in acoustic modeling. The dominance of GMM-HMMs in acoustic modeling has, over time, led to an entire "ecosystem" of front-end processing and speaker adaptation techniques specifically tailored to maximize the recognition performance under this model. Because of this, the status quo was hard to challenge with competing acoustic modeling approaches until very recently. The success of using a deep neural network acoustic model in LVCSR was first reported in [69]. In [70], the authors further presented a 33% relative improvement in WER over a discriminatively trained GMM-HMM on a 300-h English conversational telephone speech transcription task. The moniker "deep" comes from using more than one hidden layer, typically three to five. The network models the context-dependent output distributions directly and uses a greedy, layer-wise pretraining of the weights with either a supervised or unsupervised criterion [69]. This pretraining step, popularized by Hinton [71] in the context of deep belief networks (DBNs) [72], prevents the supervised training of the network from being trapped in a poor local optimum. A thorough discussion about deep NNs and their application to acoustic modeling on a variety of LVCSR tasks can be found in the article by G. Hinton et al., also included in this issue of *IEEE Signal Processing Magazine* [73].

LANGUAGE MODELING

A statistical language model (LM) $p_\Gamma(W)$ with n-gram parameters Γ represents the prior probability of a word string $W = \{w_1, \ldots, w_T\} \triangleq w_1^T$, 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 n-gram probabilities can be calculated according to the maximum likelihood estimate

$$p_{\mathrm{ML}}(w_i|w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^{i})}{\sum_{w_i} c(w_{i-n+1}^{i})}, \tag{12}$$

where $c(\cdot)$ is the count of a word sequence. 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. Although n-gram language models 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 for LVCSR. Figure 5 summarizes some new language modeling methods



[FIG5] **Overview of language modeling methods.**

that have been popular for LVCSR systems. These methods are categorized according to three issues in language modeling.

KNESER-NEY LANGUAGE MODEL

Chen [74] surveyed a series of smoothing techniques of the n-gram language model that 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 as [74], [75]

$$p_{\mathrm{KN}}(w_i|w_{i-n+1}^{i-1}) = \frac{\max\{c(w_{i-n+1}^{i}) - d_{|w_{i-n+1}^{i-1}|}, 0\}}{c(w_{i-n+1}^{i-1}\cdot)} + \frac{d_{|w_{i-n+1}^{i-1}|} N_{1+}(w_{i-n+1}^{i-1}\cdot)}{c(w_{i-n+1}^{i-1}\cdot)} p_{\mathrm{KN}}(w_i|w_{i-n+2}^{i-1}), \quad (13)$$

where $c(w_{i-n+1}^{i-1}\cdot) = \sum_u c(w_{i-n+1}^{i-1}u)$ and $N_{1+}(c_{i-n+1}^{i-1}u) = |\{u{:}c(w_{i-n+1}^{i-1}u) > 0\}|$ denotes the number of words following w_{i-n+1}^{i-1} that have one or more counts. The discount parameter $d_{|\cdot|}$ in (13) depends on the length of context w_{i-n+1}^{i-1}. This IKN language model was derived by involving marginal constraints. In addition, a modified KN (MKN) language model [74] was proposed by extending IKN language model via allowing three different discount parameters $d_{|\cdot|}$ for n grams with one $N_1(w_{i-n+1}^{i-1}\cdot)$, two $N_2(w_{i-n+1}^{i-1}\cdot)$, and three or more counts $N_{3+}(w_{i-n+1}^{i-1}\cdot)$. The MKN language model outperformed IKN language model in [74].

HIERARCHICAL PITMAN-YOR LANGUAGE MODEL

The KN language model (KNLM) was further generalized to a hierarchical Pitman-Yor (PY) language model (HPYLM) [76], where a nonparametric prior based on 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. A PY process is a generalization of a Dirichlet process with an additional discount parameter $d_{|\cdot|}$, which acts as discounting for language model smoothing. A PY process can be described by the Chinese restaurant metaphor of having an infinite number of tables, each with infinite seating capacity. Each customer (i.e., word token w_i) enters the restaurant and sits at an occupied table with probability proportional to the number of customers already sitting there, or at a new unoccupied table with probability determined by the current number of occupied tables. This PY process produces a power-law distribution that is well suited to model word frequencies in natural language [76]. HPYLM is constructed via Bayesian nonparametric learning and allows the number of n-gram parameters to grow indefinitely with large n and increasing amount of training data. HPYLM was derived as the predictive probability of a new customer w_i given the seating arrangement of w_{i-n+1}^{i-1} through collecting the probabilities of choosing occupied tables and an unoccupied table, which are labeled by w_i [76], [77]

$$p_{\mathrm{HPY}}(w_i|w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^{i}\cdot) - d_{|w_{i-n+1}^{i-1}|} N_{1+}(w_{i-n+1}^{i})}{\theta_{|w_{i-n+1}^{i-1}|} + c(w_{i-n+1}^{i-1}\cdot\cdot)} + \frac{\theta_{|w_{i-n+1}^{i-1}|} + d_{|w_{i-n+1}^{i-1}|} N_{1+}(w_{i-n+1}^{i-1}\cdot)}{\theta_{|w_{i-n+1}^{i-1}|} + c(w_{i-n+1}^{i-1}\cdot\cdot)} \times p_{\mathrm{HPY}}(w_i|w_{i-n+2}^{i-1}), \quad (14)$$

where $c(w_{i-n+1}^{i-1}\cdot\cdot) = \sum_u \sum_v c(w_{i-n+1}^{i-1}uv)$ and $\theta_{|\cdot|}$ denotes a strength parameter of a PY process depending on the length of context w_{i-n+1}^{i-1}. Gibbs sampling can be applied for model inference. Comparing (13) and (14), HPYLM is reduced to KNLM when $\theta_{|\cdot|} = 0$ and $N_{1+}(w_{i-n+1}^{i}) = 1$. HPYLM is a Bayesian generalization of KNLM with an additional strength parameter $\theta_{|\cdot|}$. In [77], HPYLM had improved performance over KNLM for LVCSR based on several large-scale training data sets.

LATENT DIRICHLET ALLOCATION LANGUAGE MODEL

To compensate the inadequate handling of long-distance information in n-gram models, latent semantic information of words and documents was explored and incorporated into the construction of large-span language models. The semantic information was represented in a low-dimensional vector space consisting of common latent topics [78]. The word clusters and the document clusters were found and used to measure the closeness between words and documents in latent semantic space. The latent semantic analysis (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 [78].

However, LSA is not a probabilistic framework and cannot be generalized for unseen test data. In [79], a topic-based language model based on probabilistic LSA (PLSALM) was proposed. PLSA parameters were estimated by ML through the EM algorithm. To tackle the generalization issue in PLSA, Blei [80] presented latent Dirichlet allocation (LDA), where Dirichlet priors were introduced to represent topic mixtures for seen documents as well as unseen documents. In [81], LDA was employed to adapt language models according to maximum a posteriori (MAP) method. The transcriptions from speech recognition were treated as a "document" to compute an LDA-adapted marginal for language model adaptation. Instead of a document-based LDA language model (LDALM) [81], a history-based LDA language model was presented to calculate the n-gram probability by [82]

$$p_{\mathrm{LDA}}(w_i|w_{i-n+1}^{i-1},\mathrm{A},\beta)=\sum_{c=1}^{C}\beta_{ic}\,\frac{g(w_{i-n+1}^{i-1},\mathrm{a}_c)}{\sum_{j=1}^{C}g(w_{i-n+1}^{i-1},\mathrm{a}_j)}. \quad (15)$$

In (15), the sequence of history words w_{i-n+1}^{i-1} is transformed to topic space or class space via a function $g(\cdot)$ by using the class-based parameter $\mathrm{A}=\{\mathrm{a}_c\}_{c=1}^{C}$. This transformation is used to find class-dependent hyperparameters of Dirichlet priors that draw the classes for a predicted word w_i. A class mixture model is established by integrating C class distributions $\beta=\{\beta_{ic}\}_{c=1}^{C}$ 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 interpolated with IKNLM and was further extended to a cache DCLM by combining the class information outside n-gram context w_{i-n+1}^{i}. This cache DCLM outperformed class-based LM [83], PLSALM, and LDALM for speech recognition [82].

MAXIMUM ENTROPY MODEL

The maximum entropy (ME) approach aims to completely model what is known, and carefully avoid assuming anything that is not known. The merit of ME model is the feasibility of merging nonindependent, asynchronous, and overlapping features into a probability model. Rosenfeld [84] proposed an ME approach to integrate diverse knowledge sources in a single unified language model. The sources of low-order n-gram, high-order n-gram, long-distance information, and syntactic/semantic knowledge are used as constraints to be imposed in an ME language model (MELM). The issues of inadequate training data and long-distance information can be addressed. Assuming that there are F features $\{f_k(\cdot)\}$ induced by the words preceding word w_i in the corresponding sentence $W^{r,i}$, the ME principle is used to estimate the MELM with ME, randomness, or smoothness while all feature functions are constrained. MELM is expressed as a log-linear model

$$p_{\mathrm{ME}}(w_i|w_{i-n+1}^{i-1},\lambda)=\frac{\exp\left[\sum_{k=1}^{F}\lambda_k f_k(W^{r,i})\right]}{\sum_{w_j}\exp\left[\sum_{k=1}^{F}\lambda_k f_k(W^{r,j})\right]}, \quad (16)$$

where $\lambda=\{\lambda_k\}$ are Lagrange multipliers that arise from a constrained optimization problem. This ME technique acts as a model smoothing method over different backoff models. The constraints due to individual features $\{f_k(\cdot)\}_{k=1}^{F}$ are imposed to equalize the true expectation $E_p[f_k]$ and the empirical expectation $E_{\tilde{p}}[f_k]$ calculated from training sentences $W=\{W\}_{r=1}^{R}$. In [84], ME parameters $\lambda=\{\lambda_k\}$ were calculated by a generalized iterative scaling procedure. The MELM was constructed by selecting long-distance trigger pairs $\{w_a\rightarrow w_b\}$ with rich mutual information from the word sequence $\{w_1,\ldots,w_a,\ldots,w_b,\ldots,w_T\}$ and treating them as feature functions for estimating ME parameters. In [85], we further mined for information-theoretic association patterns containing more than two distant words. The association pattern language model was established based on the ME framework. In [86], the ME model was extended to a joint acoustic and language model where the acoustic features extracted from HMM parameters and the linguistic features extracted from n-gram parameters were unified for joint optimization. This hybrid model characterized mutual dependencies between acoustic and linguistic features.

MODEL "M"

The ME model in (16) is known as an exponential n-gram model. Chen [87] 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 $H_{\mathrm{test}}\approx H_{\mathrm{train}}+(\gamma/N_n)\sum_{k=1}^{F}|\tilde{\lambda}_k|$, 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 [87]. This heuristic is important to explain why adding backoff n-gram features can shrink the model size and improve the model generalization. In [87], 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. The trigram based on model "M" is expressed by

$$p_{\mathrm{M}}(w_i|w_{i-2}^{i-1},\tilde{\lambda})=\sum_{c_i}p_{\mathrm{ME}}(c_i|w_{i-2}^{i-1},\tilde{\lambda})\,p_{\mathrm{ME}}(w_i|w_{i-2}^{i-1},c_i,\tilde{\lambda}), \quad (17)$$

where c_i denotes the class of word w_i. In (17), a class-based trigram based on exponential models is calculated. New features $f_{w_{i-2}w_{i-1}c_i}(w_{i-2}^{i})=\sum_{w_i\in c_i}f_{w_{i-2}w_{i-1}w_i}(w_{i-2}^{i})$ are introduced to shrink the word trigrams that differ only in w_i since trigram events $\{w_{i-2}^{i}\}$ that differ only in w_i (belonging to the same class) should have similar $\tilde{\lambda}_k$. To shrink trigram features that differ only in their histories, the author creates new features $f_{c_{i-2}c_{i-1}c_i}(w_{i-2}^{i})=\sum_{w_{i-2}\in c_{i-2},w_{i-1}\in c_{i-1}}f_{w_{i-2}w_{i-1}c_i}(w_{i-2}^{i})$. The resulting model complexity

$\sum_{k=1}^{F} |\tilde{\lambda}_k|$ is simplified. Model "M" has been successfully applied in IBM systems that were fielded in LVCSR evaluations and obtained good performance [31], [36].

NEURAL NETWORK AND SYNTACTIC LANGUAGE MODELS

N-gram language models suffer from an exponential increase in the number of parameters with the length of the word history. In contrast, an NNLM has the potential of modeling long-span dependencies with a smaller number of parameters. The application of NNs to language modeling is not straightforward because one has to transform what is inherently a discrete problem (i.e., counting words) into a continuous representation. Such a representation and corresponding NNLM has been proposed in [88] and adapted to LVCSR in [89]. The idea is to have a low-dimensional continuous feature vector for each word in the n-gram history as input to the NN, which predicts the probabilities for the most frequent words in the vocabulary. The NNLM is typically interpolated with a standard n-gram LM, which provides backoff probabilities for the words that are not modeled by the network. A recent development is to use a recurrent NNLM which, by nature, is not restricted to a word history of fixed length [90].

Another tack in language modeling research is to use syntax information to better predict words. Let's say we want to predict the word "hits" in the following sentence: "Three taken to hospital after flight from Tampa to Houston hits turbulence." The 4-gram history "Tampa to Houston" is a poor predictor of "hits" whereas "after flight" should be more effective. The structured language model (SLM) [91] uses headwords as context features that are obtained from a left-to-right parse of the sentence before the predicted word.

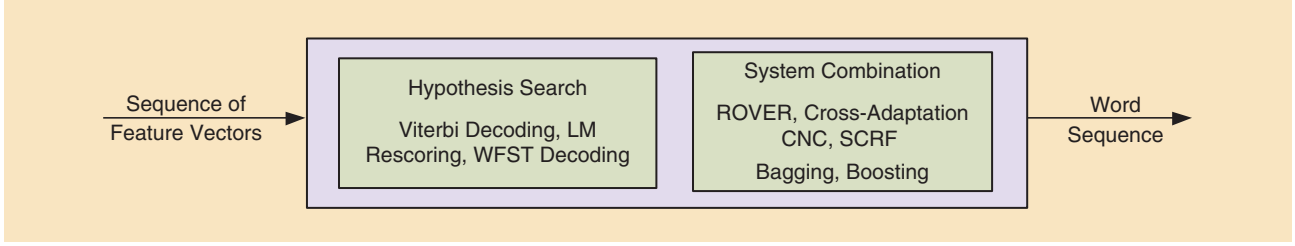
NNLM and SLM can be combined by extracting syntactic features from a parse tree and feeding their continuous representation into an NNLM. This has been done in an N-best rescoring framework for Arabic LVCSR in [92].

HYPOTHESIS SEARCH AND SYSTEM COMBINATION

Several new methods are described for hypothesis search and system combination as shown in Figure 6. These methods have a significant impact on LVCSR performance.

HYPOTHESIS SEARCH

The role of the decoder is to compute the optimal sequence of words $\hat{W}$ given the sequence of acoustic feature vectors X by incorporating information from the acoustic model and the language model via the Bayes' decision rule $\hat{W} = \arg\max_W p_\Lambda(X|W)p_\Gamma(W)$. A survey of early LVCSR decoders can be found in [14]. Since then, advances in decoding algorithms coupled with the availability of increased computing power has made accurate, real-time LVCSR possible for various domains such as broadcast news transcription or conversational telephone speech recognition [93]. 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 [94]. Such decoders have been shown to yield excellent performance when compared to classic approaches [95], [96]. One such example of a WFST decoder [45] 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) through an efficient incremental procedure described in [97]. The main advantage of using static graphs is that the graphs can be heavily optimized at "compile" time (e.g., through determinization and minimization [94]) in advance, so that minimal decoding work is required at "decode" time. However, the composition and optimization of such search networks becomes computationally challenging when large components are used. For example, it takes about 35 h to compile a search network for an Arabic LVCSR system with a vocabulary of 2.5 million words. Also, the size of the language model often exceeds the compilation limits of the search network, and the language model needs to be pruned. 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. These drawbacks of static network decoders have led us to revisit dynamic network decoders where the language model is applied dynamically. The advantages of decoupling the language model from the search network are that search network construction is much faster and the full language model can be applied directly, without the need of a rescoring pass [96].



[FIG6] Overview of hypothesis search and system combination methods.

SYSTEM COMBINATION

To obtain high levels of performance for particular domains, modern LVCSR systems employ multiple decoding and rescoring passes with several speaker adaptation passes in between. Improved system performance can be obtained by "cross-pollinating" diverse acoustic models through cross-adaptation and system combination. In contrast to self-adaptation, cross-adaptation means that the output of one system is used to adapt the acoustic models of another system. Another form of system combination pioneered by recognizer output voting error reduction (ROVER) [98] consists in aligning the word hypotheses from the different systems and in outputting the words which have the most votes within each bin. In confusion network combination (CNC) [99], consensus lattices (or "sausages") [100] from different systems are aligned and the output is given by the words with the highest posterior within each bin. In yet another approach, the lattices from multiple systems are intersected using WFST operations [94], [101]. Finally, segmental conditional random fields (SCRFs) [102] are a recent framework for combining heterogeneous acoustic models (HMM based, template based, etc.) based on exponential modeling and conditional maximum likelihood (of the word sequence given the acoustics).

The acoustic models that are combined usually differ in one or more design parameters such as input features, acoustic modeling paradigm, phonetic context, and discriminative training criterion, to name a few. Unfortunately, a lot of human intervention is required in choosing which systems are good for combination, knowledge that is often task dependent and cannot be easily transferred to other domains. Ideally, one would want an automatic procedure for training accurate systems or models that 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 [103]. Another approach is to iteratively train a sequence of acoustic models on reweighted 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 [104].

[TABLE 1] WERs AND CHARACTER ERROR RATES (CERs) FOR DIFFERENT DECODING PASSES ON ENGLISH CTS (RT04 TEST SET), ARABIC BN TRANSCRIPTION (GALE PHASE FOUR TEST SET), AND MANDARIN BN TRANSCRIPTION (GALE PHASE 2.5 TEST SET).

DECODING PASS	WER (ENGLISH CTS)	WER (ARABIC BN)	CER (MANDARIN BN)
SPEAKER INDEPENDENT	26.7%	16.7%	15.6%
SPEAKER ADAPTED	16.4%	8.9%	7.3%
LM RESCORING	15.2%	7.8%	6.5%
SYSTEM COMBINATION	—	7.4%	6.2%

Finally, in Table 1 we indicate the accuracies for the various decoding passes of three different LVCSR systems for: English conversational telephone speech (CTS) (RT04 test set) [1], Arabic broadcast news (BN) transcription (GALE Phase 4 test set) [31], and Mandarin BN transcription (GALE Phase 2.5 test set) [11]. As can be seen, substantial gains in accuracy can be obtained from speaker adaptation, language model rescoring with more sophisticated LMs (large n-gram LMs, topic-adapted LMs, model "M" and syntactic and neural network LMs [31]), and system combination using ROVER.

FUTURE DIRECTIONS

Up to now, we have described a series of approaches pertaining to all four LVCSR components that are fundamental for developing competitive LVCSR systems. In what follows, we present some new methods and point out possible directions for future LVCSR research.

STRUCTURAL STATE MODELS

HMMs with state-dependent GMMs in (1) are prevalent for LVCSR. 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 [105] 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. According to SGMMs, the state observation distribution of feature vector $\mathbf{x}_t$ at state i is expressed by a mixture of substate distributions each with a mixture of GMMs of the form

$$p_{\text{SGMM}}(\mathbf{x}_t|\Lambda_i) = \sum_{j=1}^{N_i} c_{ij} \left[\sum_{k=1}^{K} \omega_{ijk} N(\mathbf{x}_t; \mu_{ijk}, \Sigma_k) \right]. \tag{18}$$

In (18), each GMM consists of state dependent and substate dependent mixture weights $\omega_{ijk} = \exp(\mathbf{w}_k^T \mathbf{v}_{ij}) / \sum_{k'=1}^{K} \exp(\mathbf{w}_{k'}^T \mathbf{v}_{ij})$, 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 substates for state i with each substate having its own mixture weight c_{ij} and subspace vector $\mathbf{v}_{ij}$. SGMMs have two groups of parameters $\Lambda_{\text{SGMM}} = \{\Lambda_{ij}, \Lambda_k\} = \{\{c_{ij}, \mathbf{v}_{ij}\}, \{\Phi_k, \mathbf{w}_k, \Sigma_k\}\}$ that are estimated with maximum likelihood. Compared to HMM parameters $\Lambda_{\text{HMM}} = \{\omega_{ik}, \mu_{ik}, \Sigma_{ik}\}$, a much more compact representation is obtained by SGMMs due to the canonical parameters $\{\Phi_k, \mathbf{w}_k, \Sigma_k\}$ globally shared across the different states i and substates j.

SGMMs were further generalized to canonical state models (CSMs) [106] where two sets of model parameters are involved in the state likelihood calculation: the context-dependent transform parameters Λ_{ij}; and the CSM parameters Λ_k. The context-dependent state parameters are a transformed version of one or more canonical state parameters at the global level or at the context-independent phone level. The transformed parameters represent the substate parameters of a Markov state. Compared to HMMs, the structural state models using SGMMs and CSMs make the

state-space representation of acoustic features more compact and more efficient. According to CSMs, the state likelihood of $\mathbf{x}_t$ given a context-dependent state i is similar to (18) except that the mixture weights, mean vectors, and covariance matrices of the GMM are replaced by general transformation functions $\omega_{ijk} = F_\omega(k, \theta_{ij})$, $\mu_{ijk} = F_\mu(k, \theta_{ij})$, and $\Sigma_{ijk} = F_\Sigma(k, \theta_{ij})$, 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 have particular realizations such as mixtures of MLLR transforms and mixtures of fMLLR transforms and SGMMs, which differ in the transformations $F_\omega(\cdot)$, $F_\mu(\cdot)$, and $F_\Sigma(\cdot)$ that are applied to map the canonical state k to the context-dependent state i [106]. For example, the state likelihood function in case of the mixture of MLLR transforms can be realized from CSM as follows:

$$p_{\text{CSM-MLLR}}(\mathbf{x}_t|\Lambda_i) = \sum_{j=1}^{N_i} c_{ij} \left[\sum_{k=1}^{K} \omega_k N(\mathbf{x}_t; M_{ij}\xi_k, \Sigma_k) \right], \tag{19}$$

where $\xi_k = [\mu_k^T \; 1]^T$ and $\mu_{ijk} = F_\mu(k, M_{ij}) = M_{ij}\xi_k$. This CSM-MLLR model is established by transforming a canonical GMM or universal background model $\Lambda_k = \{\omega_k, \mu_k, \Sigma_k\}$ by using transform priors and context-dependent regression matrices and $\Lambda = \{c_{ij}, M_{ij}\}$. SGMMs and CSMs have been successfully applied to several LVCSR tasks [105], [106].

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 (several hundred thousand components is usually the norm) is trained to build context-dependent phone models. GMMs with Gaussian mean vectors and diagonal covariance matrices may not be an accurate representation of high-dimensional acoustic features. Alternatively, acoustic feature vectors $\mathbf{x}$ can be viewed as lying in a vector space spanned by a set of basis vectors. Such a basis representation has been popular for regression problems in machine learning and for signal recovery in the signal processing literature. This direction is now increasingly important for acoustic feature representation. For instance, in the SGMM framework, the context-dependent mean vectors $\mu_{ijk} = \Phi_k \mathbf{v}_{ij}$ in (18) are expressed via a basis representation with canonical basis vectors Φ_k and context-dependent sensing weights $\mathbf{v}_{ij}$. In [107], the full covariance matrix of a Gaussian distribution of HMMs was approximated by a set of state-independent basis vectors and state-dependent diagonal covariance matrices. The basis representation of HMM covariance matrices Σ_{ik} was effective for speech recognition. In addition, compressive sensing and sparse representation are now hot topics in the signal processing community and have been effectively exploited for speech recognition [108]. The basic idea of compressive sensing is to encode a feature vector $\mathbf{x}$ based on a set of overdetermined dictionary or basis vectors $\Phi = [\varphi_1, \ldots, \varphi_N]$ via $\mathbf{x} = \Phi\mathbf{w}$ where the sensing weights $\mathbf{w}$ are sparse and the basis vectors Φ are formed by training samples. A relatively small set of relevant basis vectors are used for sparse representation based on this exemplar-based method. The optimal sparse solution to $\mathbf{w}$ can be derived by maximizing an approximate l_1-regularized objective function [108]. However, the capability of modeling continuous speech was limited since HMMs were not integrated in the model for sparse representation of sequence data. Implementing such a memory-based method is time-consuming with high memory cost.

Consequently, Bayesian sensing HMMs (BS-HMMs) [109] were developed by incorporating Markov chains into the basis representation of continuous speech. A new Bayesian sensing framework was built for LVCSR. The underlying aspect of BS-HMMs is to measure an observed feature vector $\mathbf{x}_t$ of a speech sentence $X = \{\mathbf{x}_t\}_{t=1}^T$ based on a compact set of state-dependent dictionary $\Phi_i = [\varphi_{i1}, \ldots, \varphi_{iN}]$. The reconstruction error between measurement X and its representation $\Phi_i \mathbf{w}_t$, where $\mathbf{w}_t = [w_{t1}, \ldots, w_{tN}]^T$, is assumed to be Gaussian distributed with zero mean and a state-dependent covariance matrix or inverse precision matrix R_i^{-1}. The state likelihood function with time-dependent sensing weights $\mathbf{w}_t$ is defined by $p_{\text{BSHMM}}(\mathbf{x}_t|\mathbf{w}_t, \Phi_i, R_i) = N(\mathbf{x}_t; \Phi_i\mathbf{w}_t, R_i^{-1})$. The Bayesian perspective in BS-HMMs has its origin from the relevance vector machine (RVM) [110] which is known as a sparse Bayesian learning approach for regression and classification tasks. The purpose of Bayesian learning in BS-HMMs is to yield "distribution estimates" of the speech feature vectors due to the variations of sensing weights w_t. A Gaussian prior with zero mean and state-dependent diagonal covariance matrix is introduced to characterize the weight vector, i.e., $p(\mathbf{w}_t|A_i) = N(\mathbf{w}_t; 0, \text{diag}\{\alpha_{\text{in}}^{-1}\})$. This prior is prone to be sparse [110]. The automatic relevance determination (ARD) parameters $\{\alpha_{\text{in}}\}$ are likely to be large to draw zero values for $\mathbf{w}_t$. Only relevant basis vectors are selected to represent sequence data. Considering this Bayesian basis representation, BS-HMM parameters are formed by $\Lambda_{\text{BSHMM}} = \{\Phi_i, A_i, R_i\}$ consisting of basis vectors Φ_i and precision matrices of sensing weights A_i and reconstruction errors R_i. The predictive state likelihood function of $\mathbf{x}_t$ at state i is derived by marginalizing over sensing weights to yield [109], [111]

$$\begin{aligned} p_{\text{BSHMM}}(\mathbf{x}_t|\Lambda_i) &= \int_{R^N} p_{\text{BSHMM}}(\mathbf{x}_t|\mathbf{w}_t, \Phi_i, R_i) p(\mathbf{w}_t|A_i) d\mathbf{w}_t \\ &= N(\mathbf{x}_t; 0, (R_i - R_i\Phi_i(\Phi_i^T R_i \Phi_i + A_i)^{-1}\Phi_i^T R_i)^{-1}) \\ &= N(x_t; 0, R_i^{-1} + \Phi_i A_i^{-1}\Phi_i^T). \end{aligned} \tag{20}$$

For diagonal R_i, the state likelihood in (20) is seen as a new Gaussian distribution with a factor analyzed covariance matrix $R_i^{-1} + \Phi_i A_i^{-1}\Phi_i^T$ where the factor loading matrix $\Phi_i A_i^{-1/2}$ is seen as a rank-N correction to R_i^{-1} [111]. By applying the EM algorithm, the Type II ML estimates of BS-HMM parameters Λ_{BSHMM} are consistently formulated as implicit solutions with an efficient implementation and good convergence properties. Another important property of BS-HMMs is the ARD parameters $\{\alpha_{in}\}$ that provide model complexity control. One can initially train a large model and then prune it to a smaller size by removing basis elements that correspond to the larger ARD values [111].

Different from conventional basis representation where basis vectors and sensing weights are found separately [108], [110], BS-HMMs provide a multivariate Bayesian approach to jointly estimate the compact basis vectors and the precision matrices of sensing weights under a consistent objective function. No training examples are stored for memory-based implementation. In [109] and [112], there were several extensions for acoustic modeling: mixture models, nonzero means, speaker adaptation, and discriminative training. BS-HMMs were extended by incorporating the mixture component weights and the mean vectors in the calculation of the state likelihood, i.e., $\Lambda_{\text{BSHMM}} = \{\omega_{ik}, \mu_{ik}, \Phi_{ik}, A_{ik}, R_{ik}\}$. ML estimates of mixture weights ω_{ik} and mean vectors μ_{ik} can be easily obtained. Borrowing ideas from MLLR speaker adaptation, BS-HMMs were extended by adapting basis vectors to different speakers based on ML regression matrices. The implicit solution to regression matrices $M = \{M_c\}$ are described in [113]. Furthermore, BS-HMMs were extended from a generative model based on ML estimates to a discriminative model based on MMI estimates [112]. Model-space and feature-space BMMI training was developed and was significantly better than ML training. In the latest DARPA GALE Arabic broadcast news transcription evaluation, BS-HMMs trained on 1,800 h of data outperformed state-of-the-art HMMs with diagonal covariance GMMs even after feature-space and model-space discriminative training [111].

MODEL REGULARIZATION

ML acoustic models and language models in LVCSR systems may suffer from an overtraining problem where the estimated models are too complex to generalize for future data [51]. This leads to a limited prediction capability on unknown test sentences. In general, context-dependent GMMs with many Gaussian components are trained from a large collection of training utterances. The trained ML parameters are forced to represent the underlying distribution of the speech features given the phonetic states. But, the real-world continuous speech is collected from heterogeneous environments with mismatched training and test conditions and various sources of variations due to noise, channel, gender, speaker, accent, coarticulation, speaking rate, and emotion. 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, otherwise stated, 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 systems should tackle model regularization and compensate for the uncertainties and weaknesses in the construction of component models.

There have been several approaches developed to handle model regularization for LVCSR. Model-space and feature-space speaker adaptation [58] provides a solution to regularize the trained model parameters by compensating for the mismatch between training and test data. The language model based on model "M" [87] and the acoustic model based on BS-HMMs [113] are two new trends towards high-performance LVCSR as far as model regularization is concerned. Model "M" shrinks the exponential model to a "middle" size resulting in improved prediction performance of test data. BS-HMMs propose a Bayesian basis representation where the uncertainty of sensing weights is taken into account. A predictive Gaussian distribution with a factor analyzed covariance matrix is employed to characterize continuous speech. Nevertheless, there are other LVCSR processing components that have not been thoroughly investigated from the perspective of model regularization. For example, the issue of model selection could be extensively considered. Similar to model "M" for language modeling, the selection of middle-sized acoustic models is helpful to improve generalization for test speech. Bayesian approaches [113]–[115] have been proposed to deal with model regularization for HMMs. The common theme of these approaches is to express the uncertainties of HMM parameters by using conjugate priors. The closed-form evidence functions or predictive distributions are derived as an objective function for model complexity optimization. The evidence framework [51] has been employed to control complexity of decision trees for context-dependent acoustic modeling [114], [116] and to learn hyperparameters of HMMs for noisy speech recognition [115]. In the implementation, the prior distributions are estimated from training data and then applied to calculate the predictive distributions of test data. The uncertainty decoding [26] is performed to improve robustness in speech recognition. Also, the sparse representation provides a solution to the ill-posed problem for different models. Although BS-HMMs were motivated by RVM which performed sparse Bayesian learning, the regularization could be further enhanced by introducing a truly sparse prior and conducting Bayesian compressive sensing for LVCSR.

Previous methods require that the number of latent variables in the acoustic model and language model is fixed in advance. This is a serious limitation. Adaptively selecting the number of latent variables is an alternative direction toward achieving model regularization and improving future LVCSR. For example, in HMM-based acoustic modeling, we may need to adaptively decide the number of states for different context-dependent phonetic units and the number of Gaussian components for different context-dependent Markov states. In language modeling, we may need to choose the number of topics or classes for different n-grams. A new paradigm called Bayesian nonparametrics (BNP) [117], [118] provides an elegant solution to model complexity optimization with the least model assumption about the underlying dynamics in the data. BNP has been extensively developed for document representation and information retrieval. BNP allows the data to drive the complexity of the inferred model. The HPY language model [76] was constructed by using BNP where the number of n-gram parameters was allowed to grow indefinitely with large n. According to the paradigm of BNP, the latent variables and their number in a mixture model are automatically inferred from training data through the hierarchical Dirichlet process, which can be realized by a stick-breaking or Chinese restaurant process. A

nonparametric prior for the number of mixture components is introduced to capture the latent structure in a set of grouped data. As the amount of training data increases, the number of latent variables may grow infinitely for each group. The approximate posterior inference based on Gibbs sampling is used to implement BNP methods [117]. Furthermore, moving beyond temporal segmentation based on discrete Markov chains in HMMs, the Markov switching process [118] was developed as a more complicated process that can realize different BNP models. This process was designed to capture the continuous dynamics of multivariate time series signals. Considering the phenomena of high coarticulation and complex variations in continuous speech, the extensions of Markov switching process for acoustic modeling and language modeling will be impacting the future of LVCSR systems.

CONCLUSIONS

We have surveyed a series of approaches to front-end processing, acoustic modeling, language modeling, and back-end search and system combination that have made big contributions for LVCSR systems in the past decade or so. In the area of front-end processing, feature transformations using LDA and STC, speaker-adaptive features using VTLN and fMLLR, and discriminative features using fMPE worked well for acoustic feature extraction. In the area of acoustic modeling, feature-space and model-space discriminative training based on boosted MMI and feature-space and model-space speaker adaptation based on fMLLR and MLLR achieved the best recognition results among different methods. Alternatively, deep neural networks hold a lot of promise for acoustic modeling although training time on large amounts of data is a limiting factor. In the area of language modeling, backoff smoothing using HPYLM, large-span modeling using ME, model regularization based on model "M," and syntactic and neural network language models obtained competitive performance. Finally, in the area of hypothesis search, dynamic and WFST Viterbi decoding and system combination using ROVER, cross-adaptation, and boosting obtained good LVCSR performance. In addition, we presented flexible acoustic models based on structural state models (SGMMs and CSMs) and robust basis representation based on BS-HMMs. With the aim of modeling unknown variations in the data and model parameters, we pointed out possible future directions for LVCSR research towards model regularization for the different components of an LVCSR system.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contribution to this work of present and former members of the IBM Advanced LVCSR group: Stephen Chu, Brian Kingsbury, Jeff Kuo, Lidia Mangu, Michael Picheny, Dan Povey, Hagen Soltau, and Geoffrey Zweig.

AUTHORS

George Saon (gsaon@us.ibm.com) received his M.Sc. and Ph.D. degrees in computer science from the Henri Poincare University in Nancy, France, in 1994 and 1997. Since 1998, he has been with the IBM T.J. Watson Research Center, where he has worked on a variety of problems in the area of LVCSR. He has been a key member of IBM's speech recognition team since 2001, which participated in several U.S. government-sponsored LVCSR evaluations. He has published over 80 conference and journal papers and serves currently as an elected member of the IEEE Speech and Language Processing Technical Committee.

Jen-Tzung Chien (jtchien@nctu.edu.tw) received his Ph.D. degree from the National Tsing Hua University, Taiwan, in 1997. From 1997 to 2012, he was with the National Cheng Kung University, Taiwan. He joined the National Chiao Tung University, Taiwan, in 2012, where he is a professor in the Department of Electrical and Computer Engineering. He was a visiting professor at IBM Research in 2010. He was an associate editor of *IEEE Signal Processing Letters* from 2008 to 2011 as well as the tutorial speaker of the 2012 International Conference on Acoustics, Speech, and Signal Processing. He received the Distinguished Research Award from the National Science Council, Taiwan, in 2006 and 2010.

REFERENCES

[1] S. Chen, B. Kingsbury, L. Mangu, D. Povey, G. Saon, H. Soltau, and G. Zweig, "Advances in speech transcription at IBM under the DARPA EARS program," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 5, pp. 1596–1608, 2006.

[2] G. Evermann, H. Y. Chan, M. J. F. Gales, T. Hain, X. Liu, D. Mrva, L. Wang, and P. C. Woodland, "Development of the 2003 CU-HTK conversational telephone speech transcription system," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2004, pp. 249–252.

[3] S. Matsoukas, J.-L. Gauvain, G. Adda, T. Colthurst, C. Kao, O. Kimball, L. Lamel, F. Lefevre, J. Ma, J. Makhoul, L. Nguyen, R. Prasad, R. Schwartz, H. Schwenk, and B. Xiang, "Advances in transcription of broadcast news and conversational telephone speech within the combined EARS BBN/LIMSI system," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 5, pp. 1541–1556, 2006.

[4] A. Stolcke, B. Chen, H. Franco, V. Gadde, M. Graciarena, M. Hwang, A. Mandal, N. Morgan, X. Lei, T. Ng, M. Ostendorf, K. Sonmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, and Q. Zhu, "Recent innovations in speech-to-text transcription at SRI/ICSI/UW," *IEEE Trans. Audio Speech Lang. Processing*, vol. 14, no. 4, pp. 1729–1744, 2006.

[5] M. J. F. Gales, D. Y. Kim, P. C. Woodland, H. Y. Chan, D. Mrva, R. Sinha, and S. E. Tranter, "Progress in the CU-HTK broadcast news transcription system," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 5, pp. 1513–1525, 2006.

[6] L. Lamel, A. Messaoudi, and J.-L. Gauvain, "Improved acoustic modeling for transcribing Arabic broadcast data," in *Proc. Annu. Conf. Int. Speech Communication Association (INTERSPEECH)*, 2007, pp. 2077–2080.

[7] T. Ng, K. Nguyen, R. Zbib, and L. Nguyen, "Improved morphological decomposition for Arabic broadcast news transcription," in *Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP)*, 2009, pp. 4309–4312.

[8] M. Noamany, T. Schaaf, and T. Schultz, "Advances in the CMU/InterACT Arabic GALE transcription system," in *Proc. North American Chapter of the Association for Computational Linguistics—Human Language Technologies (NAACL-HLT)*, 2007, pp. 129–132.

[9] H. Soltau, G. Saon, B. Kingsbury, H.-K. Kuo, L. Mangu, D. Povey, and A. Emami, "Advances in arabic speech transcription at IBM under the DARPA GALE program," *IEEE Trans. Audio Speech Lang. Processing*, vol. 17, no. 5, pp. 884–894, 2009.

[10] D. Vergyri, A. Mandal, W. Wang, A. Stolcke, J. Zheng, M. Graciarena, D. Rybach, C. Gollan, R. Schlueter, K. Kirchhoff, A. Faria, and N. Morgan, "Development of the SRI/Nightingale Arabic ASR system," in *Proc. Annu. Conf. Int. Speech Communication Association (INTERSPEECH)*, 2008, pp. 1437–1440.

[11] S. M. Chu, D. Povey, H.-K. Kuo, L. Mangu, S. Zhang, Q. Shi, and Y. Qin, "The 2009 IBM GALE Mandarin broadcast transcription system," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 4374–4377.

[12] C. Plahl, B. Hoffmeister, G. Heigold, J. Loof, R. Schlueter, and H. Ney, "Development of the GALE 2008 Mandarin LVCSR system," in *Proc. Annu. Conf. Int. Speech Communication Association (INTERSPEECH)*, 2009, pp. 2307–2311.

[13] R. Sinha, M. J. F. Gales, D. Y. Kim, X. Liu, K. C. Sim, and P. C. Woodland, "The CU-HTK Mandarin broadcast news transcription system," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2006, pp. 14–19.

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

[15] G. Zweig and M. Picheny, "Advances in large vocabulary continuous speech recognition," *Adv. Comput.*, vol. 60, pp. 249–291, 2004.

[16] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, and D. O'Shaughnessy, "Developments and directions in speech recognition and understanding—Part 1," *IEEE Signal Processing Mag.*, vol. 26, no. 3, pp. 75–80, 2009.

[17] J. M. Baker, L. Deng, S. Khudanpur, C.-H. Lee, J. R. Glass, N. Morgan, and D. O'Shaughnessy, "Updated minds report on speech recognition and understanding—Part 2," *IEEE Signal Processing Mag.*, vol. 26, no. 4, pp. 78–85, 2009.

[18] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[19] S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[20] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP)*, 2000, pp. 1129–1132.

[21] M. J. F. Gales, "Maximum likelihood multiple subspace projections for hidden Markov models," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 2, pp. 37–47, 2002.

[22] 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.

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

[24] S. Furui, "Recent advances in robust speech recognition," in *Proc. Workshop Robust Speech Recognition for Unknown Communication Channels*, 1997, pp. 11–20.

[25] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, 2000, pp. 806–809.

[26] H. Liao and M. J. F. Gales, "Issues with uncertainty decoding for noise robust automatic speech recognition," *Speech Commun.*, vol. 50, no. 4, pp. 265–277, 2008.

[27] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust large vocabulary speech recognition," *IEEE Trans. Audio Speech Lang. Processing*, vol. 14, no. 3, pp. 845–854, 2006.

[28] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 1, pp. 49–60, 1998.

[29] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP)*, 1996, pp. 339–341.

[30] G. Saon, S. Dharanipragada, and D. Povey, "Feature space Gaussianization," in *Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP)*, 2004, pp. 329–332.

[31] L. Mangu, H.-K. Kuo, S. Chu, B. Kingsbury, G. Saon, H. Soltau, and F. Biadsy, "The IBM 2011 GALE Arabic speech transcription system," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011, pp. 272–277.

[32] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP)*, 2005, pp. 961–964.

[33] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 4057–4060.

[34] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP)*, 2000, pp. 1635–1638.

[35] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottleneck features for LVCSR of meetings," in *Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP)*, 2007, pp. 757–760.

[36] B. Kingsbury, H. Soltau, G. Saon, S. Chu, H. K. Kuo, L. Mangu, S. Ravuri, N. Morgan, and A. Janin, "The IBM 2009 GALE Arabic speech transcription system," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 4672–4675.

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

[38] S. Axelrod, R. Gopinath, and P. Olsen, "Modeling with a subspace constraint on inverse covariance matrices," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, 2002, pp. 2177–2180.

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

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

[41] 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. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP)*, 1986, pp. 49–52.

[42] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Comput. Speech Lang.*, vol. 16, no. 1, pp. 115–129, 2002.

[43] W. Reichl and G. Ruske, "Discriminative training for continuous speech recognition," in *Proc. European Conf. Speech Communication and Technology (EUROSPEECH)*, 1995, pp. 537–540.

[44] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP)*, 2002, pp. 105–108.

[45] G. Saon, D. Povey, and G. Zweig, "Anatomy of an extremely fast LVCSR decoder," in *Proc. Annu. Conf. Int. Speech Communication Association (INTERSPEECH)*, 2005, pp. 549–552.

[46] J. Zheng and A. Stolcke, "Improved discriminative training using phone lattices," in *Proc. Annu. Conf. Int. Speech Communication Association (INTERSPEECH)*, 2005, pp. 2125–2128.

[47] B. Zhang and S. Matsoukas and R. Schwartz, "Discriminatively trained region dependent feature transforms for speech recognition," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2006, pp. 313–316.

[48] G. Saon and D. Povey, "Penalty function maximization for large margin HMM training," in *Proc. Annu. Conf. Int. Speech Communication Association (INTERSPEECH)*, 2008, pp. 920–923.

[49] G. Saon, D. Povey, and H. Soltau, "Large margin semi-tied covariance transforms for discriminative training," in *Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP)*, 2009, pp. 3753–3756.

[50] F. Sha and L. K. Saul, "Comparison of large margin training to other discriminative methods for phonetic recognition by hidden Markov models," in *Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP)*, 2007, pp. 313–316.

[51] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.

[52] J. Li, M. Yuan, and C.-H. Lee, "Approximate test risk bound minimization through soft margin estimation," *IEEE Trans. Audio Speech Lang. Processing*, vol. 15, no. 8, pp. 2393–2404, 2007.

[53] H. Jiang, X. Li, and C. Liu, "Large margin hidden Markov models for speech recognition," *IEEE Trans. Audio Speech Lang. Processing*, vol. 14, no. 5, pp. 1584–1595, 2006.

[54].J.-C. Chen and J.-T. Chien, "Bayesian large margin hidden Markov models for speech recognition," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 3765–3768.

[55] X. He, L. Deng, and W. Chou, "Discriminative learning in sequential pattern recognition—A unifying review for optimization-oriented speech recognition," *IEEE Signal Processing Mag.*, vol. 25, no. 5, pp. 14–36, 2008.

[56] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, 1995.

[57] M. Padmanabhan, G. Saon, and G. Zweig, "Lattice-based unsupervised MLLR for speaker adaptation," in *Proc. ITRI ASR2000: ASR Challenges for the New Millennium*, 2000, pp. 128–132.

[58] D. Povey and G. Saon, "Feature and model space speaker adaptation with full covariance Gaussians," in *Proc. Annu. Conf. Int. Speech Communication Association (INTERSPEECH)*, 2006, pp. 1145–1148.

[59] G. Saon, H. Soltau, U. Chaudhari, S. Chu, B. Kingsbury, H.-K. Kuo, L. Mangu, and D. Povey, "The IBM 2008 GALE Arabic speech transcription system," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 4378–4381.

[60] H. Soltau, G. Saon, and B. Kingsbury, "The IBM Attila speech recognition toolkit," in *Proc. IEEE Workshop Spoken Language Technology (SLT)*, 2010, pp. 97–102.

[61] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Audio Speech Lang. Processing*, vol. 8, no. 4, pp. 695–707, 2000.

[62] K. Chen, W. Liau, H. Wang, and L.-S. Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, 2000, pp. 742–745.

[63] J.-T. Chien and C.-H. Huang, "Aggregate *a posteriori* linear regression adaptation," *IEEE Trans. Audio Speech Lang. Processing*, vol. 14, no. 3, pp. 797–807, 2006.

[64] A. Gunawardana and W. Byrne, "Discriminative speaker adaptation with conditional maximum likelihood linear regression," in *Proc. European Conf. Speech Communication and Technology (EUROSPEECH)*, 2001, pp. 1203–1206.

[65] L. Wang and P. C. Woodland, "MPE-based discriminative linear transforms for speaker adaptation," *Comput. Speech Lang.*, vol. 22, no. 3, pp. 256–272, 2008.

[66] J. Wu and Q. Huo, "A study of minimum classification error (MCE) linear regression for supervised adaptation of MCE-trained continuous-density hidden

Markov models," *IEEE Trans. Audio Speech Lang. Processing*, vol. 15, no. 2, pp. 478–488, 2007.

[67] M. J. F. Gales, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Processing*, vol. 4, no. 5, pp. 352–359, 1996.

[68] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, 2000, pp. 869–872.

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

[70] F. Seide, G. Li, X. Chen, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. Annu. Conf. Int. Speech Communication Association (INTERSPEECH)*, 2011, pp. 437–440.

[71] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[72] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio Speech Lang. Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[73] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Mag.* vol. 29, no. 6, pp. 82–97, 2012.

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

[75] R. Kneser and H. Ney, "Improved backing-off for *m*-gram language modeling," in *Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP)*, 1995, pp. 181–184.

[76] Y. W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes," in *Proc. Annu. Meeting of the Association for Computational Linguistics*, 2006, pp. 985–992.

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

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

[79] D. Gildea and T. Hofmann, "Topic-based language models using EM," in *Proc. European Conf. Speech Communication and Technology (EUROSPEECH)*, 1999, pp. 2167–2170.

[80] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. 1, pp. 993–1022, 2003.

[81] Y. C. Tam and T. Schultz, "Dynamic language model adaptation using variational Bayes inference," in *Proc. Annu. Conf. Int. Speech Communication Association (INTERSPEECH)*, 2005, pp. 5–8.

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

[83] P. Brown, V. Della Pietra, P. D. Souza, J. Lai, and R. Mercer, "Class-based *n*-gram models of natural language," *Comput. Linguist.*, vol. 18, no. 4, pp. 467–179, 1992.

[84] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling," *Comput. Speech Lang.*, vol. 10, no. 3, pp. 187–228, 1996.

[85] J.-T. Chien, "Association pattern language modeling," *IEEE Trans. Audio Speech Lang. Processing*, vol. 14, no. 5, pp. 1719–1728, 2006.

[86] J.-T. Chien and C.-H. Chueh, "Joint acoustic and language modeling for speech recognition," *Speech Commun.*, vol. 52, no. 3, pp. 223–235, 2010.

[87] S. F. Chen, "Shrinking exponential language models," in *Proc. North American Chapter of the Association for Computational Linguistics—Human Language Technologies (NAACL-HLT),* 2009, pp. 468–476.

[88] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, no. 2, pp. 1137–1155, 2003.

[89] H. Schwenk, "Continuous space language models," *Comput. Speech Lang.*, vol. 21, no. 3, pp. 492–518, 2007.

[90] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Annu. Conf. Int. Speech Communication Association (INTERSPEECH)*, 2010, pp. 1045–1048.

[91] C. Chelba and F. Jelinek, "Structured language modeling," *Comput. Speech Lang.*, vol. 14, no. 4, pp. 283–332, 2000.

[92] H.-K. Kuo, L. Mangu, A. Emami, I. Zitouni, and Y.-S. Lee, "Syntactic features for Arabic speech recognition," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2009, pp. 327–332.

[93] G. Saon, G. Zweig, B. Kingsbury, L. Mangu, and U. Chaudhari, "An architecture for rapid decoding of large vocabulary conversational speech," in *Proc. European Conf. Speech Communication and Technology (EUROSPEECH)*, 2003, pp. 1977–1980.

[94] M. Mohri, F. Perreira, and M. Riley, "Weighted finite state transducers in speech recognition," *Comput. Speech. Lang.*, vol. 16, no. 1, pp. 69–88, 2002.

[95] S. Kanthak, H. Ney, M. Riley, and M. Mohri, "A comparison of two LVR search optimization techniques," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, 2002, pp. 1309–1312.

[96] H. Soltau and G. Saon, "Dynamic network decoding revisited," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2009, pp. 276–281.

[97] S. Chen, "Compiling large-context decision trees into finite-state transducers," in *Proc. European Conf. Speech Communication and Technology (EUROSPEECH)*, 2003, pp. 1169–1172.

[98] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1997, pp. 347–354.

[99] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proc. Speech Transcription Workshop*, 2000.

[100] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Comput. Speech Lang.*, vol. 14, no. 4, pp. 373–400, 2000.

[101] A. Ljolje, D. Hindle, M. Riley, and R. Sproat, "The AT&T LVCSR 2000 system," in *Proc. NIST LVCSR Workshop*, 2000.

[102] G. Zweig and P. Nguyen, "A segmental CRF approach to large vocabulary continuous speech recognition," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2009, pp. 152–157.

[103] O. Siohan, B. Ramabhadran, and B. Kingsbury, "Constructing ensembles of ASR systems using randomized decision trees," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, pp. 197–200.

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

[105] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, A. Rastrow, R. C. Rose, P. Schwartz, and S. Thomas, "The subspace Gaussian mixture models—A structured model for speech recognition," *Comput. Speech Lang.*, vol. 25, no. 2, pp. 404–439, 2011.

[106] M. J. F. Gales and K. Yu, "Canonical state models for automatic speech recognition," in *Proc. Annu. Conf. Int. Speech Communication Association (INTERSPEECH)*, 2010, pp. 58–61.

[107] P. A. Olsen and R. A. Gopinath, "Modeling inverse covariance matrices by basis expansion," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 1, pp. 37–46, 2004.

[108] T. N. Sainath, B. Ramabhadran, M. Picheny, D. Nahamoo, and D. Kanevsky, "Exemplar-based sparse representation features: from TIMIT to LVCSR," *IEEE Trans. Audio Speech Lang. Processing*, vol. 19, no. 8, pp. 2598–2613, 2011.

[109] G. Saon and J.-T. Chien, "Bayesian sensing hidden Markov models for speech recognition," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 5056–5059.

[110] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, no. 6, pp. 211–244, 2001.

[111] G. Saon and J.-T. Chien, "Some properties of Bayesian sensing hidden Markov models," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011, pp. 65–70.

[112] G. Saon and J.-T. Chien, "Discriminative training for Bayesian sensing hidden Markov models," in *Proc. Int. Conf. Acoutics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 5316–5319.

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

[114] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 4, pp. 365–381, 2004.

[115].Y. Zhang, P. Liu, J.-T. Chien, and F. Soong, "An evidence framework for Bayesian learning of continuous-density hidden Markov models," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 3857–3860.

[116] J.-T. Chien and S. Furui, "Predictive hidden markov model selection for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, pp. 377–387, 2005.

[117] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies," *J. ACM*, vol. 57, no. 2, p. article 7, 2010.

[118] E. Fox, E. Sudderth, M. I. Jordan, and A. Willsky, "Bayesian nonparametric methods for learning Markov switching processes," *IEEE Signal Processing Mag.*, vol. 27, no. 6, pp. 43–54, 2010.

[SP]