On Adaptive Decision Rules and Decision Parameter Adaptation for Automatic Speech Recognition

CHIN-HUI LEE, FELLOW, IEEE, AND QIANG HUO, MEMBER, IEEE

Invited Paper

Recent advances in automatic speech recognition are accomplished by designing a plug-in maximum a posteriori decision rule such that the forms of the acoustic and language model distributions are specified and the parameters of the assumed distributions are estimated from a collection of speech and language training corpora. Maximum-likelihood point estimation is by far the most prevailing training method. However, due to the problems of unknown speech distributions, sparse training data, high spectral and temporal variabilities in speech, and possible mismatch between training and testing conditions, a dynamic training strategy is needed. To cope with the changing speakers and speaking conditions in real operational conditions for high-performance speech recognition, such paradigms incorporate a small amount of speaker and environment specific adaptation data into the training process. Bayesian adaptive learning is an optimal way to combine prior knowledge in an existing collection of general models with a new set of condition-specific adaptation data. In this paper, the mathematical framework for Bayesian adaptation of acoustic and language model parameters is first described. Maximum a posteriori point estimation is then developed for hidden Markov models and a number of useful parametric densities commonly used in automatic speech recognition and natural language processing. Other methods can be combined with Bayesian learning to enhance adaptation efficiency and effectiveness and, therefore, improve speech recognition performance. The same methodology and the set of Bayesian learning techniques can also be extended to other real-world pattern recognition problems.

Keywords*—*Acoustic modeling, adaptive decision rule, automatic speech recognition, Bayes' predictive classification rule, Bayes' risk consistency, Bayesian learning, conjugate density, expectation maximization, hidden Markov model, incomplete data problem, language modeling, maximum a posterior, maximum likelihood, maximum-likelihood linear regression, maximum mutual information, minimax classification rule, minimum classification error learning, minimum discrimination information, optimal Bayes' decision rule, prior density, quasi-Bayes' learning, recursive Bayesian learning, statistical decision theory.

Manuscript received January 15, 2000; revised May 12, 2000. The work of Q. Huo was supported in part by Hong Kong SAR under Project HKU7016/97E and in part by HKU CRCG under a Research Initiative Grant.

C.-H. Lee is with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA.

Q. Huo is with The University of Hong Kong, Hong Kong.

Publisher Item Identifier S 0018-9219(00)08099-3.

I. INTRODUCTION

Modern automatic speech recognition (ASR) technology (e.g., [12], [13], [78], [7], [134], [104], [81], and [36]) is based on an information theoretical view of the generation, acquisition, transmission, and perception of speech (e.g., [7]). Fig. 1 (adopted from B.-H. Juang's keynote speech in NNSP'96 [88]) shows a conceptual model for speech generation and signal capturing. Starting with a message M from a message source, a sequence of words W is formed through a *linguistic channel.* Different word sequences will sometimes convey the same message. It is then followed by an *articulatory channel*, which converts the discrete word sequence into a continuous speech signal S. Speaker effect, which accounts for a major portion of the speech variabilities including speech production difference, accent, dialect, speaking rate, etc., is added at this point. Additional speech distortion is introduced when the speech signal passes through the *acoustic channel*, which includes the speaking environment, interfering noise, and transducers used to capture the speech signal. This acoustic realization A is then passed through some *transmission channel* before it reaches a speech recognition system as an observed signal $\mathbf{X}$.

For speech understanding, we are interested in recovering the underlying message M from a given signal $\mathbf{X}$. On the other hand, for speech recognition, which is the focus of this paper, our goal is to "recognize" the word sequence W from the signal $\mathbf{X}$. This can also be considered as a *decision problem,* i.e., based on the information in $\mathbf{X}$ and the other relevant aspects of the problem, we attempt to make the best inference, in some sense, about W that is embedded in $\mathbf{X}$. To simplify our discussion, we view each possible word

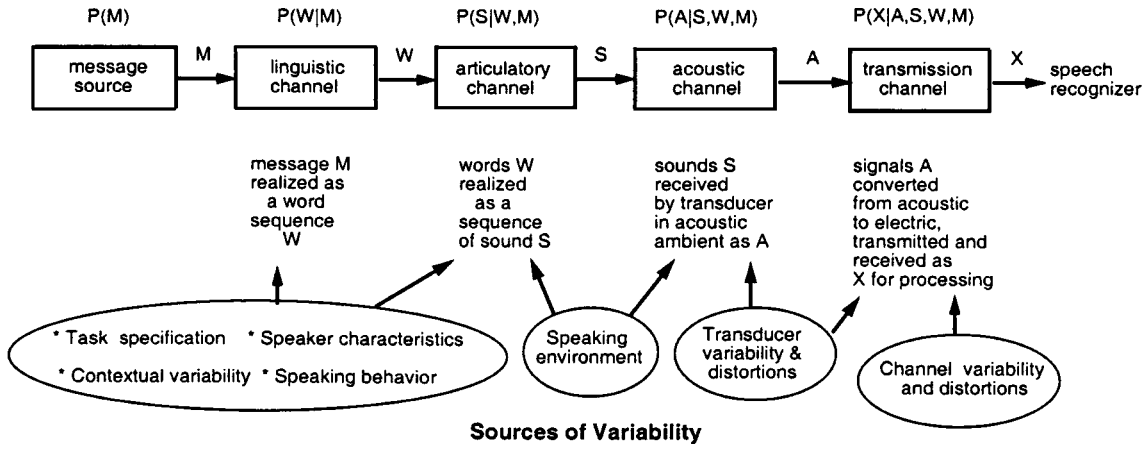

Fig. 1. Communication theoretic view of ASR.

sequence W as a *class*. Let us assume there are in total M unique classes. Therefore, speech recognition amounts to finding some optimal decision rules for classification of the observation $\mathbf{X}$ into one of M fixed classes. Depending on the evaluation criterion, there exist many decision rules. Not all of them are of equal value in practice. Because of the different sources of variability, as shown in Fig. 1, the speech signal $\mathbf{X}$ is usually featured by *uncertainty, variability, lack of determinism*, and *stochasticity*. This makes the statistical pattern-matching paradigm a natural choice for formulating and solving the ASR problem. If the joint distribution $p(W, \mathbf{X})$ is specified exactly, the *Bayes' decision rule* (e.g., [42], [142], and [95]) is implemented as follows:

$$\hat{W} = \arg\max_{W} p(W, \mathbf{X}) \tag{1}$$

with $\hat{W}$ being the recognized sentence. This decision rule is known to be optimal for minimizing the decision risks. Due to the complex channel interactions in Fig. 1, it is unlikely that we have complete knowledge to specify the joint distribution of $\mathbf{X}$ and W.

For real-world practical ASR problems, it is also difficult to characterize the individual channels in Fig. 1. A simplified *source-channel* model, as shown in Fig. 2 is usually adopted as follows.

1) The joint distribution $p(W, \mathbf{X})$ is decomposed into two components, $p(\mathbf{X}|W)$ and $P(W)$, known as an acoustic model and a language model, respectively. The former evaluates the likelihood of the observation $\mathbf{X}$ assuming the word sequence W is given, and the latter computes the language probability of W.
2) The forms of $p(\mathbf{X}|W)$ and $P(W)$ are assumed *parametric* probability density functions (pdfs), i.e., $p_{\Lambda}(\mathbf{X}|W)$ and $P_{\Gamma}(W)$, respectively.
3) The parameters Λ and Γ of the above distributions are to be estimated from some *training data* by using some particular *point estimation* techniques.

Therefore, all the contributions of the intermediate channels, such as articulatory, acoustic, and transmission channels, are lumped together as a noisy channel. Speech recognition is now solved as a *channel decoding* problem in which channel

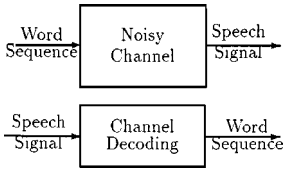

Fig. 2. Source-channel model of ASR.

modeling, which includes acoustic and language modeling from some training data, becomes a critical issue. With these simplifications, the most popular way to solve the ASR problem is to use the well-known *plug-in* maximum *a posteriori* (MAP) *decision rule* (e.g., [42], [142], and [95]):

$$\begin{aligned}\hat{W} &= \arg\max_{W} P(W|\mathbf{X}) \\ &= \arg\max_{W} p_{\hat{\Lambda}}(\mathbf{X}|W) \cdot P_{\hat{\Gamma}}(W)\end{aligned} \tag{2}$$

where $\hat{\Lambda}$ and $\hat{\Gamma}$ are the estimated parameters obtained during training and $\hat{W}$ is the recognized sentence during testing. This decision rule, derived from the optimal Bayes' decision rule, is also widely used in many other pattern-recognition applications.

To implement the plug-in MAP decision rule, there are three major research areas: 1) *search* procedures that find the optimal solution $\hat{W}$ from the large class space; *dynamic programming* (DP) and *delayed decision* approaches are commonly employed; 2) *speech feature representation*, which extracts relevant speech parameters that are easy to model, less susceptible to measurement noise and distortion in adverse conditions, and give high discrimination power; and 3) *acoustic* and *language modeling* that choose the set of units to model and the algorithms to estimate the parameters Λ and Γ. In this paper, we concentrate our discussion on the third issue of density parameter estimation.

Currently, the most widely adopted and the most successful modeling approach to ASR is to use a set of hidden Markov models (HMMs) as the acoustic models of subword or whole-word units, and to use the statistical N-gram model or its variants as language models for words and/or word classes. The readers are referred to good tutorials in [133] and [79] for an introduction to the above approaches

and their applications. By using the above-mentioned plug-in MAP decision rule, it has been repetitively shown by experiments in the past three decades that given a large amount of *representative* training speech and text data, good statistical models of speech and language can be constructed to achieve a high performance for a wide range of ASR tasks. This has given the speech research community a certain level of confidence in believing that the *discrete HMM* (DHMM, e.g., [107]), the *tied mixture* or *semicontinuous HMM* (TMHMM or SCHMM, e.g., [16] and [67]), and the mixture Gaussian *continuous density HMM* (CDHMM, e.g., [85] and [132]), together with N-gram models (e.g., [60] and [93]), provide a good approximate parametric forms for $p_\Lambda(\mathbf{X}|W)$ and $P_\Gamma(W)$, respectively. Although these models are apparently imperfect, they are mathematically well defined and capable of simultaneously modeling both the spectral and temporal variation in speech. They are also well thought of because they both fit into the framework of *finite-state* representations [111] of *knowledge sources* so that the speech-recognition problem can be solved as a *network search* problem over a complex network representation of speech and language. In addition, new models are constantly being explored (e.g., [38], [61], and [129]). Based on the belief that these acoustic and language models are good approximates, the *maximum-likelihood* (ML) estimate for the HMM parameters [14], [112], [85] and N-gram model parameters (e.g., [79]) has been the most popular parameter-estimation method.

However, due to many problems, caused by incorrect model specification and the *curse of dimension* in estimating a large number of parameters with only a limited amount of training data, there is often an observed performance degradation when using ML estimators in cross-condition testing. One major reason lies in the possible mismatch between the underlying acoustic characteristics associated with the training and testing conditions. This mismatch may arise from inter- and intraspeaker variabilities, transducer, channel, and other environmental variabilities, and many other phonetic and linguistic effects due to task mismatch. To bridge this performance gap, one possible solution is to design a speech-recognition system that is robust to the above types of acoustic mismatch, and this has been a long-standing objective of many researchers over the past 20 years. Another way to reduce the possible acoustic mismatch is to adopt the so-called *adaptive learning* approach. The scenario is like this: starting from a pretrained (e.g., *speaker* and/or *task independent* [66], [105]) speech-recognition system, for a new user (or a group of users) to use the system for a specific task, a small amount of adaptation data is collected from the user. These data are used to construct an adaptive system for the speaker in the particular environment for that specific application. By doing so, the mismatch between training and testing can generally be reduced and the speech-recognition performance is greatly enhanced.

The topic of HMM parameter estimation and adaptation is one of the most fruitful areas in the field of automatic speech recognition in recent years. Key technical advances are summarized in the roadmap shown in Fig. 3. We will come back to discuss this roadmap in more detail later. Roughly speaking, there are two major classes of adaptation techniques, namely, 1) *direct* classifier parameter adaptation, which adapts the HMM parameters through Bayesian learning, such as MAP estimation (e.g., [102], [55], [56], and [69]); and 2) *indirect* classifier parameter adaptation through estimation of some *transformation* or *structure* parameters (e.g., [49]), using ML or MAP estimation. These two types of adaptation techniques for HMMs are illustrated in Fig. 4.

Bayesian adaptive learning is an optimal way to combine prior knowledge in an existing collection of general models with a new set of condition-specific adaptation data. However, when too many parameters need to be adapted at the same time while too little adaptation data is available, one often relies on an *auxiliary* structure with less parameters to be adapted. The most often used structure is through an *affine* transformation such as finding *linear regression* transformation of the mean vectors of the original HMMs. Both *maximum-likelihood linear regression* (MLLR [109], [39]) and *maximum a posteriori linear regression* (MAPLR [154]) have been adopted with good success. *Joint Bayesian adaptation* of both HMM and transformation parameters has also been developed [155]. It is believed that further advances will be made in the area of adaptation and compensation in order to improve the *robustness* and *performance* of speech-recognition system.

In this paper, we attempt to explain, from a *statistical decision* point of view, why the pattern-recognition approach to ASR works so well in certain conditions, and more important, why it does not work as well in many other situations. The rest of this paper is organized as follows. In Section II, the rationale for using the decision theoretic approach to designing plug-in MAP decision rules for the ASR problem is presented. In Section III, general issues regarding estimation of acoustic and language model parameters are addressed. Adaptive point estimation will be justified. In Section IV, MAP estimation, which is a prevailing Bayesian learning paradigm in speech recognition, is formulated for some popular acoustic and language models. The key issues of prior density specification and hyperparameter estimation will also be discussed. Besides batch adaptation, on-line incremental adaptation is of practical importance and requires a new *recursive Bayesian learning* and prior evolution formulation, which is developed in Section V. In Section VI, some ML and Bayesian learning algorithms for structure parameters are presented. Through a small set of structure parameters, these techniques are designed to enhance learning efficiency and effectiveness, especially for cases with only sparse adaptation data but a large number of classifier parameters needs to be adapted. In Section VII, we briefly discuss the dual issue of unsupervised adaptation and decision rule compensation. Many algorithms originally developed for adaptation can be extended to compensation and vice versa. It is important to know how adaptation techniques can be used to improve robustness and compensate for performance degradation in real-world, operational ASR systems. In Section VIII, we present some recent

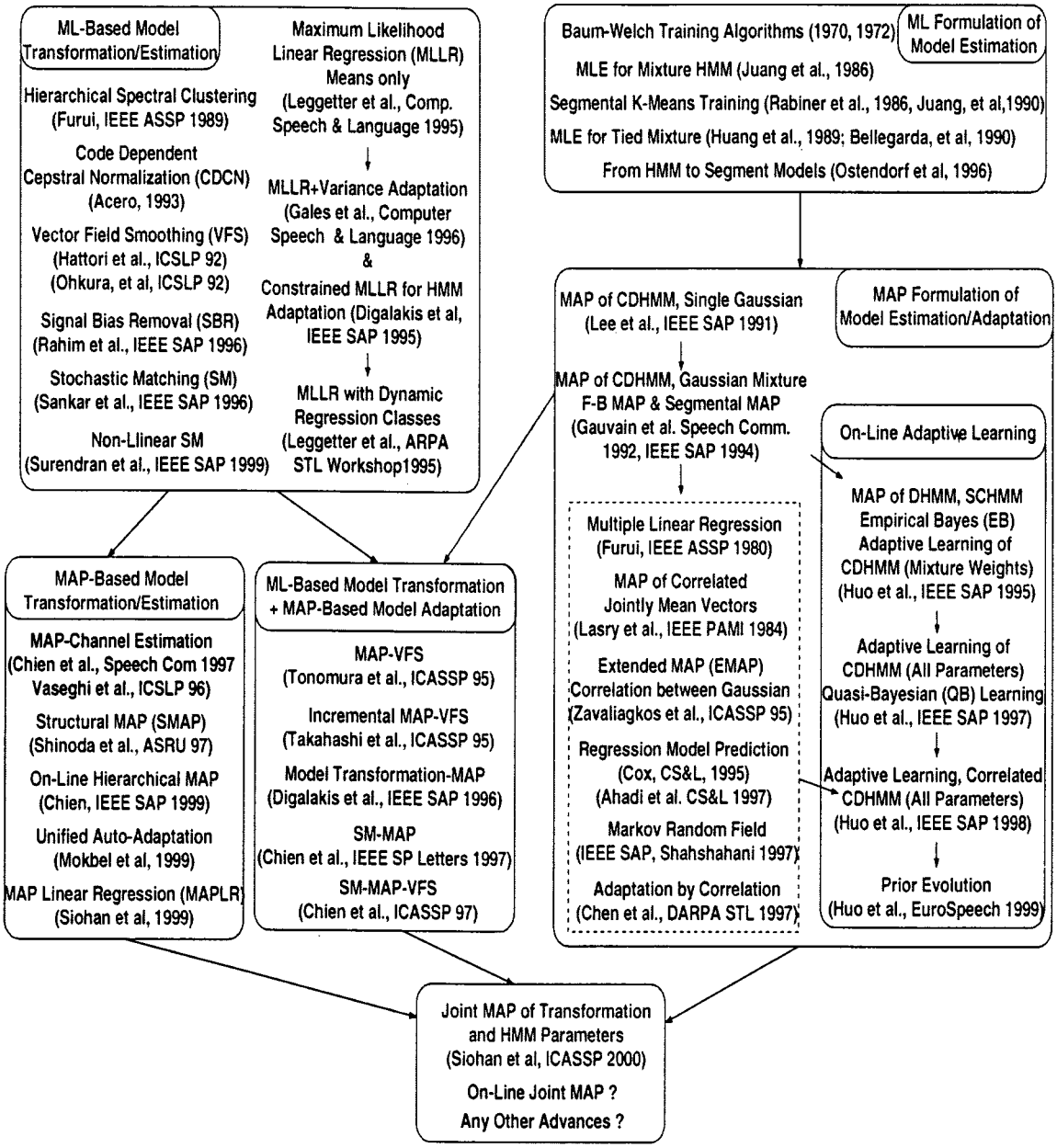

Fig. 3. Advances in estimation and adaptation of HMMs.

advances in extending plug-in MAP decision rules to new classification rules, such as *minimax classification* [122] and *Bayesian predictive classification* [74], [83]. Adaptation and compensation techniques can also be incorporated into this new class of recognition algorithms. Finally, we summarize our findings in Section IX.

II. Statistical Decision Theory

In its simplest form, let us assume that pattern recognition problem of interest is to classify a given observed signal $\mathbf{X}$ into one of M classes, $W \in \Omega_W$, where $\Omega_W = \{W_1, W_2, \ldots, W_M\}$ denotes the set of M classes. In the case of speech recognition, a class $W \in \Omega_W$ may be of any linguistic unit, e.g., a phoneme, syllable, word, phrase, semantic concept or attribute, sentence, etc. The signal $\mathbf{X}$ is usually a feature vector sequence extracted from a speech utterance. Let us assume that $\mathbf{X}$ belongs to a suitable signal space Ω_x. The pattern-recognition problem is, in principle, equivalent to finding a *decision rule* $d(\cdot)$ in a set of possible decision rules $\mathcal{D}$, such that d: $\Omega_x \rightarrow \Omega_W$, or simply

$$W = d(\mathbf{X}), \qquad \text{for } \mathbf{X} \in \Omega_x,\ W \in \Omega_W,\ \text{and } d(\cdot) \in \mathcal{D} \quad (3)$$

with W being one of the M possible class labels in Ω_W. In this case, the *decision space* $\{d(\mathbf{X})\colon \mathbf{X} \in \Omega_x\}$ of the decision rule $d(\cdot)$ is the same as the Ω_W. A decision rule $d(\cdot) \in \mathcal{D}$ implies a mapping from the sample space to the class label space. This mapping is known as a *nonrandomized decision rule* [48]. Define $\Omega_x(W_i) = \{\mathbf{X}\colon \mathbf{X} \in \Omega_x,\ d(\mathbf{X}) = W_i\}$ to be a subset of Ω_x corresponding to the region of $\mathbf{X}$ being mapped as class W_i with the decision rule $d(\cdot)$. Then the

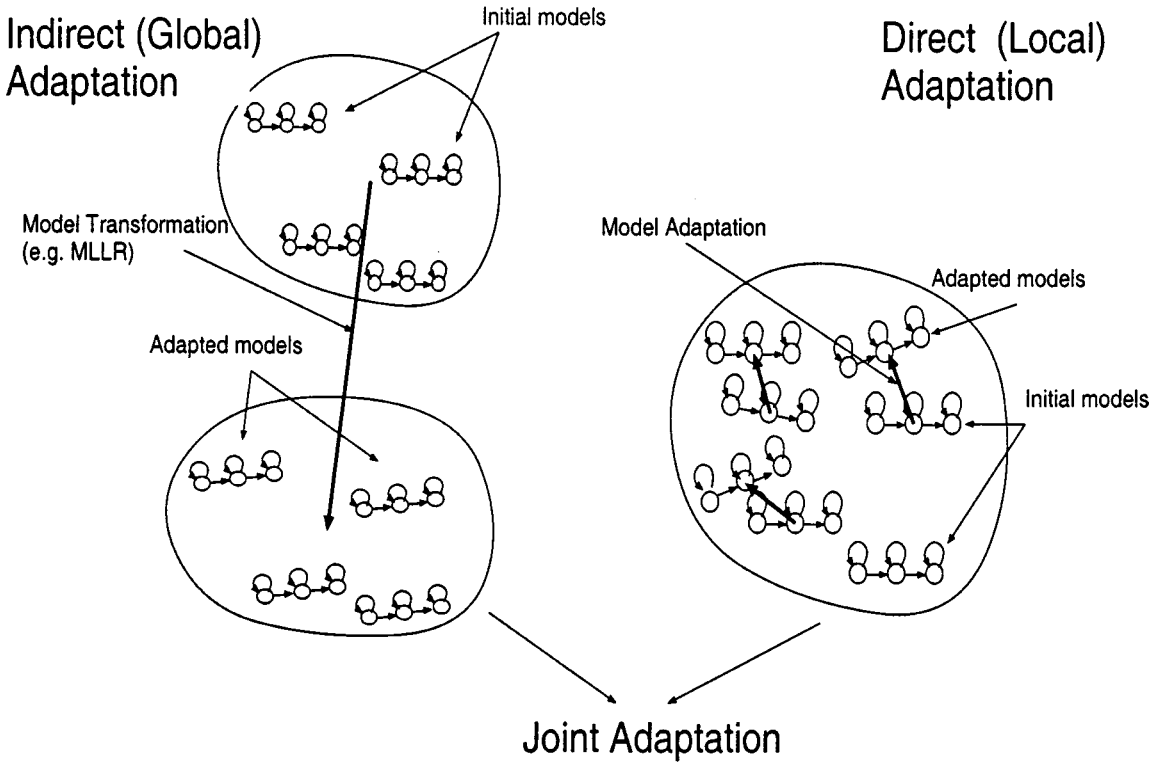

Fig. 4. Direct and indirect adaptation of HMMs.

construction of a decision rule amounts to finding a partition $\Omega_x(d(\cdot)) = \{\Omega_x(W_1), \Omega_x(W_2), \ldots, \Omega_x(W_M)\}$ of the observation space Ω_x under the following constraints:

$$\bigcup_{i=1}^{M} \Omega_x(W_i) = \Omega_x, \quad \Omega_x(W_i) \bigcap \Omega_x(W_j) = \emptyset, \quad \text{for } i \neq j;\ i, j = 1, 2, \ldots, M. \tag{4}$$

There may exist an infinite set of decision rules for the same given classification problem. Not all of them are of equal value in practice though. To determine whether a decision rule is "good," one has to agree on a reasonable set of criteria for assessing the "goodness." Let us show one possible formulation by using the classical statistical decision theory pioneered by Wald [171] and developed by many others (e.g., [48] and [42]).

A. Optimal Bayes' Decision Rule for Known Distributions

Let us view W and an observation $\mathbf{X}$ as a jointly distributed random pair $(W, \mathbf{X})$, whose joint pdf is denoted by $p(W, \mathbf{X})$. In the so-called *sampling paradigm*, we can decompose $p(W, \mathbf{X})$ into a product of the class prior probability $P(W)$ and the class conditional pdf $p(\mathbf{X}|W)$, i.e., $p(W, \mathbf{X}) = p(\mathbf{X}|W)P(W)$. One way of formalizing a goodness criterion is to use the knowledge of the possible consequences of the decisions. Often this knowledge can be quantified by assigning a *loss* that would be incurred for each possible decision. Let $\ell(W, d(\mathbf{X}))$ be the *loss function* associated with making a decision $d(\mathbf{X})$ if the true class is W. One would like the loss function to have the following property:

$$0 \leq \ell(W, W) \leq \ell(W, d(\mathbf{X}) \neq W). \tag{5}$$

If we assume the *true distribution* $p(W, \mathbf{X})$ is known, then the conditional and marginal distributions, namely, $p(\mathbf{X}|W)$, $p(W|\mathbf{X})$, $P(W)$, and $p(\mathbf{X})$, can be calculated. Now we can define the *total risk* $r(d(\cdot))$ for a decision rule $d(\cdot)$ as an expected value of the loss function, i.e.,

$$r(d(\cdot)) = \mathbf{E}_{W,\mathbf{X}}[\ell(W, d(\mathbf{X}))]$$

$$= \sum_{W \in \Omega_W} \int_{\mathbf{X} \in \Omega_x} \ell(W, d(\mathbf{X})) p(W, \mathbf{X})\, d\mathbf{X} \tag{6}$$

$$= \int_{\mathbf{X} \in \Omega_x} p(\mathbf{X}) \left[\sum_{W \in \Omega_W} \ell(W, d(\mathbf{X})) P(W|\mathbf{X}) \right] d\mathbf{X} \tag{7}$$

$$= \sum_{W \in \Omega_W} P(W) \int_{\mathbf{X} \in \Omega_x} \ell(W, d(\mathbf{X})) p(\mathbf{X}|W)\, d\mathbf{X} \tag{8}$$

where $\mathbf{E}_{W,\mathbf{X}}[\cdot]$ denotes mathematical expectation with respect to the distribution of $(W, \mathbf{X})$. The above total risk can be used as a measure of the quality of decision rules. Usually

the less the total risk, the better is the decision rule. In this framework, the issue of constructing an optimal decision rule becomes the following risk minimization problem:

$$\min_{d(\cdot)\in\mathcal{D}} r(d(\cdot)) = \min_{d(\cdot)\in\mathcal{D}} \int_{\mathbf{X}\in\Omega_x} p(\mathbf{X}) \left[\sum_{W\in\Omega_W} \ell(W, d(\mathbf{X})) P(W|\mathbf{X}) \right] d\mathbf{X}. \tag{9}$$

This optimization can be solved by minimizing the expression in the square brackets in the above equation. It is clear that the solution leads to the following optimal decision rule:

$$d_o(\mathbf{X}) = \arg \min_{d(\mathbf{X})\in\Omega_W} \sum_{W\in\Omega_W} \ell(W, d(\mathbf{X})) P(W|\mathbf{X}) \tag{10}$$

which is also known as the *Bayes' decision rule.* The resulting minimum total risk

$$r(d_o(\cdot)) = \int_{\mathbf{X}\in\Omega_x} p(\mathbf{X}) \left[\sum_{W\in\Omega_W} \ell(W, d_o(\mathbf{X})) P(W|\mathbf{X}) \right] d\mathbf{X} \tag{11}$$

is called the *Bayes' risk.* This risk value is the best that can be achieved if the distribution $p(W, \mathbf{X})$ is known.

In speech recognition, a reasonable option is to assume that every misclassification of $\mathbf{X}$ is equally serious, thereby resulting in the so-called *0–1 loss function*

$$\ell(W, d(\mathbf{X})) = \begin{cases} 0, & \text{if } W = d(\mathbf{X}) \text{(correct decision)} \\ 1, & \text{if } W \neq d(\mathbf{X}) \text{(wrong decision)} \end{cases} \tag{12}$$

for $W \in \Omega_W$, $d(\mathbf{X}) \in \Omega_W$. Substituting (12) into (8), we obtain

$$r_{01}(d(\cdot)) = \sum_{W\in\Omega_W} P(W) \int_{\mathbf{X}\notin\Omega_x(W)} p(\mathbf{X}|W)\, d\mathbf{X} \tag{13}$$

$$= 1 - \sum_{W\in\Omega_W} \int_{\mathbf{X}\in\Omega_x(W)} P(W) p(\mathbf{X}|W)\, d\mathbf{X}. \tag{14}$$

Therefore, in the case of the 0–1 loss function, the total risk is the unconditional error probability, which is apparently a good measure of the quality of decision rules for the ASR task. The optimal decision rule $d_{01}(\cdot)$ under the *minimum classification error (MCE)* criterion with the 0–1 loss function is then solved as $d_{\mathrm{MAP}}(\mathbf{X}) = \hat{W}$ such that

$$\hat{W} = \arg\max_W P(W|\mathbf{X}) = \arg\max_W p(\mathbf{X}|W) \cdot P(W) \tag{15}$$

which is also known as the *MAP decision rule.*

In summary, in constructing these optimal decision rules, it was assumed that complete prior information about the classes is known, i.e.:

1) the observation space Ω_x is given;
2) the loss function $\ell(W, d(\mathbf{X}))$ is given;
3) the true pdf $p(W, \mathbf{X})$ or $p(\mathbf{X}|W)$ and $P(W)$ are known.

Under these assumptions, the optimality criterion is the minimization of the risk functional $r(d(\cdot))$, and the optimal decision rule is the Bayes' decision rule.

B. Plug-In Decision Rule for Unknown Distributions

In practice, we know neither the *true* parametric form of the joint distribution $p(W, \mathbf{X})$ nor its *true* parameters. We shall say that we have *prior uncertainty* [95] in this case. If we have some labeled *independent* training sample set, $\mathcal{X} = \{(W^i, \mathbf{X}^i); i = 1, 2, \ldots, n\}$, we can reduce the prior uncertainty by constructing a decision rule from $\mathcal{X}$. The decision rule $d(\cdot) = d(\mathbf{X}|\mathcal{X})$ based on the training set $\mathcal{X}$ and used to classify a random observation $\mathbf{X}$ that is *independent* of $\mathcal{X}$, is called an *adaptive decision rule* [95]. There are several principles that can be used for the construction of such rules. The most popular family of adaptive decision rules might be the so-called *plug-in decision rules.*

For this approach, let $\{\hat{P}(W), \hat{p}(\mathbf{X}|W)\}$ be any statistical estimators of the true distributions $\{P(W), p(\mathbf{X}|W)\}$ based on the training set $\mathcal{X}$. The *plug-in decision rule* [58] is the adaptive decision rule, $d = \hat{d}_o(\mathbf{X})$, derived from the Bayes' decision rule in (10) by substitution of the estimators $\{\hat{P}(W), \hat{p}(\mathbf{X}|W)\}$ for the unknown true distributions $\{P(W), p(\mathbf{X}|W)\}$

$$\hat{d}_o(\mathbf{X}) = \arg \min_{d(\mathbf{X})\in\Omega_W} \sum_{W\in\Omega_W} \ell(W, d(\mathbf{X})) \hat{P}(W|\mathbf{X}) \tag{16}$$

where

$$\hat{P}(W|\mathbf{X}) = \frac{\hat{p}(\mathbf{X}|W)\hat{P}(W)}{\sum_W \hat{p}(\mathbf{X}|W)\hat{P}(W)}. \tag{17}$$

By varying the loss function and by using different kinds of estimators $\{\hat{P}(W), \hat{p}(\mathbf{X}|W)\}$, a fairly rich family of plug-in decision rules can be obtained. For example, adopting the 0–1 loss function will lead to the following plug-in decision rule, $\hat{d}_{\mathrm{MAP}}(\mathbf{X}) = \hat{W}$, such that

$$\hat{W} = \arg\max_W \hat{P}(W|\mathbf{X}) = \arg\max_W \hat{p}(\mathbf{X}|W) \cdot \hat{P}(W) \tag{18}$$

which is also known as the *plug-in MAP decision rule.*

It can be shown [58] that the plug-in decision rule $\hat{d}_o(\cdot)$ in (16) minimizes the *plug-in risk* $\hat{r}(d(\cdot))$, which is an estimate of the total risk using the *density plug-in estimator* $\{\hat{P}(W), \hat{p}(\mathbf{X}|W)\}$, i.e.,

$$\hat{d}_o(\cdot) = \arg \min_{d(\cdot)\in\mathcal{D}} \hat{r}(d(\cdot)) \tag{19}$$

where

$$\hat{r}(d(\cdot)) = \sum_{W\in\Omega_W} \hat{P}(W) \int_{\mathbf{X}\in\Omega_x} \ell(W, d(\mathbf{X})) \hat{p}(\mathbf{X}|W)\, d\mathbf{X}. \tag{20}$$

The minimum plug-in risk is then $\hat{r}(\hat{d}_o(\cdot))$.

C. Bayes' Risk Consistency

As noted in [58], the plug-in risk $\hat{r}(\hat{d}_o(\cdot))$ of the plug-in Bayes' decision rule in (16), is often less than its total risk

$r(\hat{d}_o(\cdot))$ and is even optimistically biased as an estimator of the Bayes' risk $r(d_o(\cdot))$.

Property: If the estimators $\{\hat{P}(W), \hat{p}(\mathbf{X}|W)\}$ are pointwise unbiased, then

$$\mathbf{E}[\hat{r}(\hat{d}_o(\cdot))] \leq r(d_o(\cdot)) \leq r(\hat{d}_o(\cdot)). \tag{21}$$

However, the usefulness of the plug-in Bayes' decision rule in (16) can be justified by the following theorem of *Bayes' risk consistency* [58].

Theorem: (Bayes' Risk Consistency): If the estimators $\{\hat{P}(W), \hat{p}(\mathbf{X}|W)\}$ are strongly consistent, i.e., converge to the true distributions almost surely as the training sample size n increases ($n \to \infty$)

$$\begin{aligned} \hat{P}(W) &\xrightarrow{a.s.} P(W) \\ \hat{p}(\mathbf{X}|W) &\xrightarrow{a.s.} p(\mathbf{X}|W), \quad \text{for } W \in \Omega_W \text{ and } \mathbf{X} \in \Omega_x \end{aligned} \tag{22}$$

then the plug-in risk for the plug-in decision rule in (16) is a strongly consistent estimator of the Bayes' risk, i.e.,

$$\hat{r}(\hat{d}_o(\cdot)) \xrightarrow{a.s.} r(d_o(\cdot)). \tag{23}$$

D. Violation of Modeling Assumptions

The principles of the construction of the above-mentioned optimal decision rule and plug-in decision rules are based on some assumptions that may be violated in practice. From the computational modeling point of view, there are three main distortion types that produce violations of assumptions summarized as follows [95]:

1) distortions caused by small-sample effects;
2) distortions of models for training samples;
3) distortions of models for observations to be classified.

The distortions caused by small-sample effects are typical for all statistical plug-in procedures. They arise from the noncoincidence of the statistical estimates $\{\hat{P}(W), \hat{p}(\mathbf{X}|W)\}$ of probability characteristics and their true values $\{P(W), p(\mathbf{X}|W)\}$. We want to emphasize again that the plug-in decision rules described in the previous section are asymptotically optimal only when:

1) the training samples $\mathcal{X} = \{(W^i, \mathbf{X}^i); i = 1, \ldots, n\}$ are collected by a series of *independent* experiments such that $(W^i, \mathbf{X}^i) \sim p(W, \mathbf{X})$, or more intuitively speaking, $\mathcal{X}$ should be *representative* enough with respect to the true distribution of the testing data $\mathbf{X}$;
2) training sample size $n \to \infty$, i.e., there is sufficient amount of training data available.

In practice, the training sample set $\mathcal{X}$ always has a finite size (i.e., $n < \infty$), and in many cases, is possibly also not representative enough. The random deviations of statistical estimators, $(\hat{P}(W) - P(W), \hat{p}(\mathbf{X}|W) - p(\mathbf{X}|W))$, can then produce significant increases of the decision risk. So, the design and/or collection of the training samples become very critical. The key is to make the samples in $\mathcal{X}$ follow the intended distribution $p(W, \mathbf{X})$ as closely as possible. Otherwise, some more intelligent ways of using the available training data must be developed.

As for the distortions of the models for the training samples, they can be caused by the wrong assumptions and/or inflexible parametric forms of the model, the mislabeling of training samples, outliers in training samples, etc. To cope with these problems, better models and techniques need to be developed for robust learning from data.

The biggest problem for ASR might be caused by the third type of distortion, the distortions of the models for the observations to be classified. In most real applications, there always exists some form of mismatch, which causes a distortion between the trained models and the test data. These mismatches, some of them identified in Fig. 1, may arise from inter- and intraspeaker variabilities; transducer, channel, and other environmental variabilities; and many other phonetic and linguistic effects due to the problem of task mismatch. How to achieve the performance robustness in this context has become one of the most active research areas in ASR in the past decade.

III. Parametric Models and Parameter Estimation

As we mentioned above, because of the constraints of the limited computational resources and training data in practical ASR applications, we always have to *assume* some parametric form for $p(W, \mathbf{X})$, e.g., via $p_\Lambda(\mathbf{X}|W)$ and $P_\Gamma(W)$. The parameter set (Λ, Γ) has to be *estimated* from a given training set $\mathcal{X}$ by using certain parameter estimation techniques. The above Bayes' risk consistency theorem tells us that it is often possible to construct plug-in procedures that are *Bayes' risk consistent* in the sense that the sequence of plug-in risks converges to the Bayes' risk as the training sets increase in size. However, there is an important assumption behind this argument, that is, the assumed distributions $p_\Lambda(\mathbf{X}|W)$ and $P_\Gamma(W)$ obey the parametric structure in question. In order to achieve a good approximation to reality, some flexible parametric models should be adopted.

A. Point Estimation of Decision Rule Parameters

As we pointed out in the introduction section, so far, the most successful modeling approach to ASR is to use a set of HMMs as the acoustic models of subword or whole-word units and to use the statistical N-gram model or its variants as language models for words and/or word classes. Based on the belief that these acoustic and language models are good approximates, the widespread use of the plug-in MAP decision rule with the ML estimators can be justified by using the above Bayes' risk consistency theorem due to the following facts:

1) the ML estimator of (Λ, Γ) is strongly consistent, unbiased, and efficient;
2) this can then be translated into strong distribution consistency if the chosen parametric forms of $p_\Lambda(\mathbf{X}|W)$ and $P_\Gamma(W)$ are indeed correct.

According to our knowledge, it was Nadas [125] who first provided such an insight for the speech recognition community.

The topic of ML estimation of HMM parameters have been developed extensively in the last two decades (e.g., [14],

[112], and [85]). The readers are referred to some excellent tutorials (e.g., [133]) and textbooks (e.g., [47], [134], and [81]) for the formulation. An *HMM tool kit* (HTK) with source codes and software routines is also available (e.g., [174] and [179]). Of course, one can always argue that although the ML estimators $\hat{\Lambda}$ and $\hat{\Gamma}$ may be excellent estimators of Λ and Γ, there is no guarantee that $P_{\hat{\Gamma}}(W)$ and $p_{\hat{\Lambda}}(\mathbf{X}|W)$ are good guesses for $P(W)$ and $p(\mathbf{X}|W)$. Nor $\hat{d}_o(\cdot)$ is necessarily a good approximation to $d_o(\cdot)$. The performance of the plug-in rules and other procedures should really be tied to the classification accuracy instead of the behavior of $(\hat{\Lambda}, \hat{\Gamma})$ as a *point estimator* for (Λ, Γ). This has motivated many studies in the past two decades aiming at a good alternative to ML training. One method is *minimum discrimination information* (MDI) training [43], which adjusts the HMM parameters to minimize the *discrimination information*, or *directed divergence*, between the assumed HMM distribution and the best possible distribution derived from the training data under certain constraints embedded in the training data. Unfortunately, no significant experimental results have been reported to show how MDI works in a speech-recognition task. Another class of approaches is the so-called *discriminative training* method. Some of them, such as *maximum mutual information* (MMI) training [8], *conditional maximum-likelihood estimate* (CMLE) [127], and *H-criteria* [62], aim indirectly at reducing the error rate of the speech recognizer on the training data. Other methods, such as *corrective training* [11], *minimum empirical error rate* training [44], [115], and (MCE) training [4], [91], [87], [28], [89], [92], try to reduce the recognition error rate on training data in a more direct way. Among these approaches, MCE formulation has been the most successful, which we will examine briefly in the following.

For the MCE approach, we view a decision rule $d(\cdot)$ as a *discriminant function*. The discriminant $d(\mathbf{X})$ classifies observation $\mathbf{X}$ into one of the M classes. When MCE training is formulated as minimizing an approximate *empirical classification error* [59], [87] or *expected classification error* [4], it can be solved by using *generalized probabilistic descend* (GPD) and segmental GPD algorithms (e.g., [92], [27], and [113]). It have been extensively studied and successfully applied to speaker recognition (e.g., [113]), speech recognition (e.g., [28], [29], [138], [105], and [136]), *utterance verification* (e.g., [161] and [137], optical character recognition (e.g., [177]), and many other applications referred to in [92].

So far we have considered the following two design principles, namely, 1) plug-in MAP decision rule with ML density estimators and 2) discriminant classifier with minimum empirical/expected classification error training. The following conclusions may be made concerning these two strategies:

1) the asymptotic behavior of the first approach will depend on the appropriateness (in the sense of estimator consistency) of the parametric forms of the assumed distributions;
2) while the asymptotic behavior of the second approach will depend on the choice of the discriminant function. Theoretically speaking, it is not clear yet which strategy is better for a moderately sized training set.

There are already many studies on the second issue of discriminative training, as we have described above. In this paper, we focus our discussion on the plug-in decision strategy with a chosen form of parametric densities and the corresponding estimation techniques for designing plug-in decision rules.

B. Challenges in Speech and Language Model Estimation

In the past, most ASR systems rely on a *static* design strategy in that all the knowledge sources needed in a system, including acoustic models of speech units, lexical models of words and phrases, and language models of word sequences, are acquired at the design phase and remain the same during the testing phase. Many good studies on acoustic modeling are available in literature (e.g., [107], [101], [66], [10], [77], [178], and [174]). Equally as many papers are concerned with language modeling (e.g., [93], [9], [79], [100], [131], and [18]). The performance of the ASR systems usually depends on how close the training data cover the statistical variation of the signal and language from the training to the testing conditions and on how well the feature representation and the trained models capture the relevant information for discriminating among different speech and linguistic units. Since it is not practical to collect a large set of speech and text examples, spoken and written by a large population over all possible combinations of signal conditions, it is likely that the conditions in testing are different from those in training. Such a mismatch is a major source of error for conventional pattern-matching systems. A state-of-the-art system may perform poorly when the test data are collected under a totally different signal condition.

Regarding to the possible mismatches, both linguistic and acoustic mismatches might occur. A *linguistic mismatch* is mainly caused by incomplete task specifications, inadequate knowledge representations, and insufficient training data, etc. On the other hand, an *acoustic mismatch* between training and testing conditions arises from various sources, including difference in desired speaking formats, task specifications, and signal realizations. For example, task model and vocabulary usage heavily influence the efficacy of the training process. For a given task, speech models trained based on task-dependent data usually outperform models trained with task-independent data. Similarly, speech models trained based on isolated word data usually have problems capturing the coarticulation effect between words and, therefore, often perform not as well for continuous speech recognition. Another major source of acoustic mismatch derives from changing signal conditions. For example, changes in transducers, channels, speaking environments, speaker population, speaking rates, speaking styles, echoes, and reverberations, and the combination of them all contribute to performance degradation. In addition to the previously discussed linguistic and acoustic mismatches, *model incorrectness* and estimation error also cause robustness problems for a recognizer. Since the distortion mechanism and the

exact signal models are often unknown or only partially known, it makes such robustness problems more difficult to manage.

C. Adaptive Speech and Language Modeling

An alternative to relying the performance solely on an appropriate training set is to use a *dynamic* design strategy. Starting from an initial set of models, new information is constantly collected during the use of the system and incorporated into the system using *adaptive learning* algorithms. In this way, the set of models can be adapted over time (with new training material, possibly derived from actual test utterances) to the task, the language, the speaker, and/or the environment (e.g., [160], [102], [148], [68], [103], and [165]). Such methods of adaptive training are usable for new speakers, tasks, and environments, and will be shown later to be an effective way of creating a good set of problem-specific models (adaptive models) from a more general set of models (which are speaker, environment, task, and probably context independent). This can be accomplished, among many possibilities, by MAP estimation of HMM parameters (e.g., [102], [55], [56], [69], and [139]) or ML/MAP estimation of a small number of transformation or structure parameters (e.g., [109], [39], [147], [24], and [154]).

For adaptation of language model parameters, it involves simultaneous estimation of many probability parameters under constraints. This is still a growing area of intensive research. Such adaptive techniques include Bayesian methods (e.g., [46]), MDI-based algorithms (e.g., [34], [94], and [140]), *maximum entropy* approaches (e.g., [100] and [145]), and adaptive learning mechanisms using a *history* or *cache* (e.g., [79], [80], and [98]), or a *trigger* (e.g., [100]). In the following, we limit our discussion on adaptive acoustic modeling. It is noted that many of the principles and techniques presented here are equally applicable to adaptive language modeling (e.g., MAP in [46]). The readers are also referred to two papers discussing state-of-the-art language modeling techniques [146], [19] in this issue.

A list of recent advances of acoustic parameter estimation and adaptation is summarized in the roadmap in Fig. 3. Starting from the classical ML estimation approaches to estimating HMM parameters shown in the upper right block of the roadmap, there are a number of important developments aiming at accommodating adaptive learning of a huge number of HMM parameters, typically on the order of a few million for large-vocabulary continuous speech recognition. The first major area of work is *direct* MAP estimation of HMM parameters, which is summarized in the lower right block of the roadmap. The second major area is *indirect* ML estimation of structure parameters, which in turn provide HMM parameter estimation through some form of transformations. This is summarized in the upper left block of the roadmap. Once the framework of these two major areas is established, one could apply the MAP estimation approaches to structure parameter adaptation as shown in the left branch of the bottom left blocks of the roadmap. One could also combine direct and indirect estimation and perform hybrid ML/MAP estimation as shown in the right branch of the bottom left block of the roadmap. Finally, a formal development of joint MAP estimation of transformation and HMM parameters is recently established [155]. It serves as a unified framework to combine direct and indirect classifier parameter estimation/adaptation in designing adaptive decision rules for automatic speech recognition. It is noted that batch and on-line joint estimation of the above two sets of parameters shown in the bottom of Fig. 3 provides a natural extension for adapting classifier parameters. Many other technology convergence paradigms can also be worked out.

Although the theoretical discussion in Section II offers some directions for designing adaptive decision rules, there are some practical difficulties dealing with real-world pattern-recognition problems, such as speech recognition. In this paper, we focus our discussion on adaptation techniques that have been developed to address some of the concerns. We first formulate MAP estimation algorithms for a number of parametric densities commonly used in ASR in Section IV. Detailed development is given to illustrate the procedure for deriving MAP estimates in missing data problems such as the case of adaptive learning of HMM parameters. The key issue of prior specification, which is critical in many Bayesian learning problems, is also presented. We then establish the theory of recursive Bayesian learning through prior evolution, which is important for on-line Bayesian adaptation. These three sets of fundamentals, i.e., MAP estimation, prior evolution, and recursive adaptation, form the basis for many recent advances in Bayesian adaptive learning for ASR. We will also briefly discuss the important topics of parameter reduction, correlation interpolation, tying, and structure. They serve as useful side information to improve the efficiency and effectiveness of adaptive learning for large systems. Since there are too many HMM parameters to be estimated, adaptation through structures and constraints of the parameters is of important concern and will be discussed in Section VI. Adaptive learning techniques can also be used to improve robustness of an ASR system by adapting the system according to the testing data. This is known as *compensation* or adaptation without supervision. We will address this family of problems and point the relationship between adaptation and compensation in Section VII.

IV. MAP-Based Bayesian Adaptation

In the following discussion, we focus our attention on techniques specifically developed for direct adaptation of HMM parameters. Since point estimates are required to implement the plug-in MAP decoder in (2), we used the *Bayesian learning* principle to derive MAP estimates of the parameters of some useful acoustic and speech models. The *prior density* needed in the MAP formulation is specified based on prior knowledge embedded in a large collection of data or in a set of speech and language models. The Bayesian learning framework offers a way to incorporate newly acquired application-specific data into existing models and combine them in an optimal manner. It is, therefore, an

efficient technique for handling the sparse training data problem, which is typical in adaptive learning of model parameters. It is also noted that techniques and issues discussed here can be used to derived MAP estimates of other decision rule parameters, such as structure parameter estimation to be addressed in Section VI.

Three key issues arise in the MAP formulation, namely:

1) the definition of prior densities for the model parameters of interest;
2) the estimation of the prior density parameters, sometimes referred to as *hyperparameters;*
3) the solution to MAP estimation.

All three issues are related, and a good definition of the prior densities is crucial in resolving these issues. For acoustic modeling of speech units and language modeling of linguistic units, continuous-variable observations are often characterized by *multivariate Gaussian* densities and *gamma* densities; and discrete-variable observations are often modeled by *multinomial* distributions. For example, in hidden Markov modeling, all the above three densities from the exponential family have been combined to characterize the initial probabilities, the transition probabilities, the histogram of discrete state output probabilities for discrete HMMs, the mixture gains for tied-mixture HMMs and continuous density HMMs with mixture Gaussian state densities, the duration probability, the N-gram probabilities in language modeling, etc. In most cases, the use of the *conjugate prior* formulation, such as a *Dirichlet density* for the estimation of the parameters of multinomial pdfs and a *normal-Wishart density* for the estimation of the parameters of Gaussian pdfs, has been found effective [102], [55], [56], [69].

The MAP-based adaptive learning algorithms have been applied to a number of applications, including speaker and task adaptation [102], [55], [103], [69], context adaptation [55], corrective training [55], parameter smoothing [102], [55], speaker group modeling [55], on-line incremental adaptation with stored history data [120], and N-gram and histogram probability smoothing and adaptation [55]. The same approach can also be extended to the problems of speaker normalization, nonnative speaker adaptation, rapid speaker enrollment, transducer and channel adaptation, speaking environment adaptation, etc.

For a given set of training/adaptation data $\mathbf{X}$, the conventional ML estimation assumes that the HMM parameter λ is *fixed* but *unknown* and solves

$$\hat{\lambda}_{\mathrm{ML}} = \arg\max_{\lambda} f(\mathbf{X}|\lambda) \tag{24}$$

where $f(\mathbf{X}|\lambda)$ is the likelihood of $\mathbf{X}$. On the other hand, the MAP formulation assumes the parameter λ to be a *random* vector with a certain distribution. Furthermore, there is an assumed correlation between the observation vectors and the parameters so that a statistical inference of λ can be made using a small set of adaptation data $\mathbf{X}$. Before making any new observations, the parameter vector is assumed to have a *prior* density $g(\lambda)$. When new data $\mathbf{X}$ are incorporated, the parameter vector is characterized by a *posterior density* $g(\lambda|\mathbf{X})$. The MAP estimate maximizes the posterior density

$$\hat{\lambda}_{\mathrm{MAP}} = \arg\max_{\lambda} g(\lambda|\mathbf{X}) = \arg\max_{\lambda} f(\mathbf{X}|\lambda)g(\lambda). \tag{25}$$

Since the parameters of a prior density can, among many possibilities, also be estimated from an existing HMM λ_0, this framework provides a way to combine $\lambda_0$with newly acquired data $\mathbf{X}$ in an optimal manner.

The prior density $g(\lambda)$ characterizes statistics of the parameters of interest before any measurement was made. It can be used to impose constraints on the values of the parameters. If the parameter is fixed but unknown and is to be estimated from the data, then there is no preference to what the value of the parameter should be. In such a case, the prior distribution $g(\lambda)$ is often called a *noninformative prior,*, which is a constant for the entire parameter region of interest. The MAP estimate obtained by solving (25) is, therefore, equivalent to the ML estimate obtained by solving (24). When the prior of the HMM parameters is assumed to be the product of the conjugate priors for all HMM parameters, the MAP estimates can be solved with the *expectation–maximization* (EM) algorithm [56]. A theoretical framework of MAP estimation of HMM was first proposed by Lee *et al.* [102] for estimating the mean and the covariance matrix parameters of a CDHMM with a multivariate Gaussian state observation density. It was then extended to handle all the HMM parameters, including the initial state probabilities, the transition probabilities, the duration density probabilities, the energy histogram probabilities, and the state observation probabilities, of a CDHMM with mixture Gaussian state density [55], [56]. The same Bayesian formulation has also been applied to the estimation of the parameters of discrete HMMs and of tied-mixture (or semicontinuous) HMMs [69].

In analogy to the two well-known ML estimation approaches, the *forward–backward MAP* [56] and the *segmental MAP* [102], [55], [56] algorithms have been developed to solve for the MAP estimates. When conjugate priors for the *complete-data* densities are assumed, the MAP estimates can be expressed as a weighted sum of two components: one depends on the information in the prior density (e.g., λ_0) and the other depends on the new set of adaptation data [56]. It can further be shown that the MAP and the ML estimates are *asymptotically equivalent* [56]. We now describe MAP adaptation algorithms for some useful parametric densities commonly used in speech recognition.

A. MAP Estimation of Multinomial Densities

Let ω_k be the probability of observing the kth discrete event e_k among a set of K possible outcomes $\{e_k, k = 1, \ldots, K\}$ and $\sum_{k=1}^{K} \omega_k = 1$. Then, the probability of observing a sequence of independently and identically distributed (i.i.d.) discrete observations $\mathbf{X} = (x_1, \ldots, x_T)$ follows a multinomial distribution

$$p(x_1, \ldots, x_T|\omega_1, \ldots, \omega_K) \propto \prod_{k=1}^{K} \omega_k^{n_k} \tag{26}$$

where $n_k = \sum_{t=1}^{T} 1(x_t = e_k)$ is the number of occurrence of observing the kth event in the sequence with $1(\mathcal{L})$ being the indicator function defined on the logical variable $\mathcal{L}$. Here, we use "$\propto$" to denote proportionality. Many useful random variables used in speech recognition and language processing, including N-grams, histograms, mixture gains, and discrete HMM probabilities, can be modeled this way. The prior density of $(\omega_1, \ldots, \omega_K)$ can be assumed as a Dirichlet density (e.g., [33]), which is a conjugate prior for the parameters of a multinomial density, i.e.,

$$g(\omega_1, \ldots, \omega_K) \propto \prod_{k=1}^{K} \omega_k^{\nu_k - 1} \tag{27}$$

where $\{\nu_k > 0, k = 1, \ldots, K\}$ is the set of hyperparameters. The MAP estimate can be easily solved as (e.g., [33])

$$\tilde{\omega}_k = \frac{n_k + \nu_k - 1}{\sum_{k=1}^{K} (n_k + \nu_k - 1)}. \tag{28}$$

B. MAP Estimation of Multivariate Gaussian Mixtures

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ be a sample of T i.i.d. vector observations drawn from a mixture of K D-dimensional multivariate normal densities

$$f(\mathbf{x}|\theta) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{x}|m_k, r_k) \tag{29}$$

where

$$\theta = (\omega_1, \ldots, \omega_K, m_1, \ldots, m_K, r_1, \ldots, r_K) \tag{30}$$

is the parameter vector and ω_k denotes the mixture gain for the kth mixture component subject to the constraint $\sum_{k=1}^{K} \omega_k = 1$. $\mathcal{N}(\mathbf{x}|m_k, r_k)$ is the kth normal density function denoted by

$$\mathcal{N}(\mathbf{x}|m_k, r_k) \propto |r_k|^{1/2} \exp\left[-\tfrac{1}{2}(\mathbf{x} - m_k)^t r_k (\mathbf{x} - m_k)\right] \tag{31}$$

where m_k is the D-dimensional mean vector and r_k is the $D \times D$ precision matrix, which is defined as the inverse of the covariance matrix Σ_k, i.e., $r_k^{-1} = \Sigma_k$. Here, we use $|r|$ to denote the determinant of a matrix r and r^t to denote the transpose of the matrix or vector r. In the following, we will also use $\mathrm{tr}(r)$ to denote the trace of the matrix r.

Given $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, its joint pdf (or the likelihood function of θ) is specified by the equation[1]

$$f(\mathbf{X}|\theta) = \prod_{t=1}^{T} \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{x}_t|m_k, r_k). \tag{32}$$

It is well known that no *sufficient statistics* of a fixed dimension exists for the parameter vector θ in (30) (e.g., [141], [159], [64], and [56]), therefore, no joint conjugate prior densities can be specified. However, a finite mixture density can be interpreted as a density associated with a statistical population, which is a mixture of K component populations with mixing proportions $(\omega_1, \ldots, \omega_K)$. In other words, $f(\mathbf{X}|\theta)$ can be viewed as a marginal pdf with the parameter θ of a joint pdf expressed as the product of a multinomial density (for the sizes of the component populations) and multivariate Gaussian densities (for the component densities). If we view $(\omega_1, \ldots, \omega_K)$ as the parameter vector of a multinomial density, then the joint conjugate prior density for $(\omega_1, \ldots, \omega_K)$ is a Dirichlet density as shown before. Similarly, for the vector parameter (m_k, r_k) of the individual Gaussian mixture component, the joint conjugate prior density is a normal-Wishart density [33] of the form

$$\begin{aligned} &g(m_k, r_k|\varphi_k) \\ &\quad \propto |r_k|^{(\alpha_k - D)/2} \exp\Big[-\frac{\tau_k}{2}(m_k - \mu_k)^t r_k (m_k - \mu_k)\Big] \\ &\qquad \cdot \exp\big[-\tfrac{1}{2}\mathrm{tr}(u_k r_k)\big] \end{aligned} \tag{33}$$

where $\varphi_k = (\tau_k, \mu_k, \alpha_k, u_k)$ is the hyperparameter vector such that $\alpha_k > D-1, \tau_k > 0, \mu_k$ is a vector of dimension D and u_k is a $D \times D$ positive definite matrix. Assuming independence between the parameters of the individual mixture components and the set of the mixture weights, the joint prior density $g(\theta)$ is the product of the prior pdfs of the form

$$g(\theta) = g(\omega_1, \ldots, \omega_K) \prod_{k=1}^{K} g(m_k, r_k). \tag{34}$$

The EM algorithm is an iterative procedure for approximating ML estimates in the context of incomplete-data cases such as mixture density and hidden Markov model estimation problems [15], [37]. This procedure consists of maximizing, at each iteration the auxiliary function, $Q(\theta|\overline{\theta})$ defined as the expectation of the *complete-data* log-likelihood given the incomplete data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ and the current fit $\overline{\theta}$. For a mixture density, the complete-data likelihood is the joint likelihood of $\mathbf{X}$ and the unobserved labels referring to the mixture components, $\mathbf{l} = (l_1, \ldots, l_T)$, i.e., $\mathbf{Y} = (\mathbf{X}, \mathbf{l})$. By defining the auxiliary function as $Q(\theta|\overline{\theta}) = E[\log h(\mathbf{Y}|\theta)|\mathbf{X}, \overline{\theta}]$. The EM procedure derives from the facts that $\log f(\mathbf{X}|\theta) = Q(\theta|\overline{\theta}) - H(\theta|\overline{\theta})$ where $H(\theta|\overline{\theta}) = E[\log h(\mathbf{Y}|\mathbf{X}, \theta)|\mathbf{X}, \overline{\theta})]$ and $H(\theta|\overline{\theta}) \leq H(\overline{\theta}|\overline{\theta})$, and, therefore, whenever a value θ satisfies $Q(\theta|\overline{\theta}) > Q(\overline{\theta}|\overline{\theta})$ then $f(\mathbf{X}|\theta) > f(\mathbf{X}|\overline{\theta})$. It follows that the same iterative procedure can be used to estimate the mode of the posterior density by maximizing the auxiliary function, $R(\theta|\overline{\theta}) = Q(\theta|\overline{\theta}) + \log g(\theta)$, at each iteration instead of the maximization of $Q(\theta|\overline{\theta})$ in conventional ML procedures [37].

Let $\Psi(\theta|\overline{\theta}) = \exp R(\theta|\overline{\theta})$ be the function to be maximized. Define the following *membership* function for the mixture Gaussian density

$$c_{kt} = \frac{\overline{\omega}_k \mathcal{N}(\mathbf{x}_t|\overline{m}_k, \overline{r}_k)}{\sum_{l=1}^{K} \overline{\omega}_l \mathcal{N}(\mathbf{x}_t|\overline{m}_l, \overline{r}_l)}. \tag{35}$$

[1]In this study, the same term $f(\cdot)$ is used to denote both the joint and the marginal pdfs since it is not likely to cause confusion.

Using the equality

$$\sum_{t=1}^{T} c_{kt}(\mathbf{x}_t - m_k)^t r_k(\mathbf{x}_t - m_k) = c_k(m_k - \overline{\mathbf{x}}_k)^t r_k(m_k - \overline{\mathbf{x}}_k) + \mathrm{tr}(S_k r_k) \quad (36)$$

it follows from the definition of $f(\mathbf{X}|\theta)$ and $Q(\theta|\overline{\theta})$ that

$$\Psi(\theta|\overline{\theta}) \propto g(\theta) \prod_{k=1}^{K} \omega_k^{c_k} |r_k|^{c_k/2} \cdot \exp\Big[-\frac{c_k}{2}(m_k - \overline{\mathbf{x}}_k)^t - r_k(m_k - \overline{\mathbf{x}}_k)\tfrac{1}{2}\,\mathrm{tr}(S_k r_k)\Big] \quad (37)$$

where $c_k = \sum_{t=1}^{T} c_{kt}$, $\overline{\mathbf{x}}_k = \sum_{t=1}^{T} c_{kt}\mathbf{x}_t / c_k$, and $S_k = \sum_{t=1}^{T} c_{kt}(\mathbf{x}_t - \overline{x}_k)(\mathbf{x}_t - \overline{x}_k)^t$ are weighted count, weighted sample mean vector, and weighted sample covariance matrix for the kth mixture component.

It can easily be verified from (37) and (34) that $\Psi(\cdot|\overline{\theta})$ belongs to the same distribution family as $g(\cdot)$, and they form a conjugate pdf family for the complete-data density. The mode of $\Psi(\cdot|\overline{\theta})$, denoted by $(\hat{\omega}_k, \hat{m}_k, \hat{r}_k)$, may be obtained from the modes of the Dirichlet and normal-Wishart densities based on well-known formulation of these pdfs in statistics literature [33]. Thus, the EM reestimation formulas are derived as follows:

$$\hat{\omega}_k = \frac{(\nu_k - 1) + \sum_{t=1}^{T} c_{kt}}{\sum_{l=1}^{K}\left[(\nu_l - 1) + \sum_{t=1}^{T} c_{lt}\right]} \quad (38)$$

$$\hat{m}_k = \frac{\tau_k \mu_k + \sum_{t=1}^{T} c_{kt}\mathbf{x}_t}{\tau_k + \sum_{t=1}^{T} c_{kt}} \quad (39)$$

$$\hat{r}_k^{-1} = \frac{u_k + \sum_{t=1}^{T} c_{kt}(\mathbf{x}_t - \hat{m}_k)(\mathbf{x}_t - \hat{m}_k)^t + \tau_k(\mu_k - \hat{m}_k)(\mu_k - \hat{m}_k)^t}{(\alpha_k - D) + \sum_{t=1}^{T} c_{kt}}. \quad (40)$$

It can be seen that the new parameter estimates are simply a weighted sum of the prior parameters and the observed data (a form applicable to both parameter smoothing and adaptation). If it is assumed that $\hat{\omega}_k > 0$, then the EM reestimation formulas for the MAP and ML approaches are asymptotically equivalent [55], a desirable property in many applications. According to our knowledge, it was Hamilton who first developed the MAP estimation of parameters for mixtures of normal distributions, under the name of the quasi-Bayesian approach [64]. We want to warn the readers here that the *quasi-Bayesian* algorithm to be developed later in this paper is a completely different approach.

C. MAP Estimation of HMM Parameters

The development in the previous section for a mixture of multivariate Gaussian densities can be extended to the case of HMM with Gaussian mixture state observation densities. For notational convenience, it is assumed that the observation pdfs of all the states have the same number of mixture components.

Consider an N-state CDHMM with parameter vector $\lambda = (\pi, \mathbf{A}, \theta)$, where π is the initial probability vector, $\mathbf{A}$ is the transition probability matrix, and θ is the pdf parameter vector composed of the mixture parameters $\theta_i = [(w_{ik}, m_{ik}, r_{ik}): k = 1, \ldots, K]$ for each state i.

For a sample $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, the complete data is $\mathbf{Y} = (\mathbf{X}, \mathbf{s}, \mathbf{l})$, where $\mathbf{s} = (s_0, \ldots, s_T)$ is the unobserved state sequence, and $\mathbf{l} = (l_1, \ldots, l_T)$ is the sequence of the unobserved mixture component labels, $s_t \in [1, 2, \ldots, N]$ and $l_t \in [1, 2, \ldots, K]$. It follows that the pdf of $\mathbf{X}$ has the form

$$f(\mathbf{X}|\lambda) = \sum_{\mathbf{s}} \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} \left[\sum_{k=1}^{K} \omega_{s_t k} \mathcal{N}(\mathbf{x}_t | m_{s_t k}, r_{s_t k})\right] \quad (41)$$

where

- π_i — initial probability of state i;
- a_{ij} — transition probability from state i to state j;
- $\theta_{ik} = (m_{ik}, r_{ik})$ — parameter vector of the kth normal pdf associated with state i, and the first summation in (41) is over all possible state sequences.

If no prior knowledge is assumed about $\mathbf{A}$ and π, or alternatively if these parameters are assumed fixed and known, the prior density $g(\cdot)$ can be chosen to have the following form $g(\lambda) = \prod_i g(\theta_i)$, where $g(\theta_i)$ is defined by (34). In the general case where MAP estimation is applied not only to the observation density parameters but also to the initial and transition probabilities, the prior density for all the HMM parameters can be assumed as

$$g(\lambda) \propto \prod_{i=1}^{N}\left[\pi_i^{\eta_i - 1} g(\theta_i) \prod_{j=1}^{N} a_{ij}^{\eta_{ij} - 1}\right] \quad (42)$$

where $\{\eta_i\}$ is the set of parameters for the prior density of the initial probabilities $\{\pi_i\}$ and $\{\eta_{ij}\}$ is the set of parameters for the prior density of transition probabilities $\{a_{ij}\}$, all defined the same way as a Dirichlet density.

In Section IV-C2, we examine two ways of approximating λ_{MAP} by local maximization of $f(\mathbf{X}|\lambda)g(\lambda)$ or $f(\mathbf{X}, \mathbf{s}|\lambda)g(\lambda)$. These two solutions are the MAP versions of the forward–backward algorithm [15] and of the segmental k-means algorithm [132], [86], algorithms that were developed for ML estimation.

1) Forward–Backward MAP Estimation: In the ML formulation, the auxiliary function of the EM algorithm

can be decomposed into a sum of three auxiliary functions $Q_\pi(\pi|\overline{\lambda})$, $Q_A(\mathbf{A}, \overline{\lambda})$, and $Q_\theta(\theta, \overline{\lambda})$ such that they can be independently maximized [85]. The three functions take the following forms:

$$Q_\pi(\pi|\overline{\lambda}) = \sum_{i=1}^{N} \gamma_{i0} \log \pi_i \tag{43}$$

$$\begin{aligned} Q_A(\mathbf{A}|\overline{\lambda}) &= \sum_{i=1}^{N} Q_{a_i}(a_i|\overline{\lambda}) \\ &= \sum_{i=1}^{N} \left[\sum_{t=1}^{T} \sum_{j=1}^{N} \xi_{ijt} \log a_{ij} \right] \end{aligned} \tag{44}$$

$$\begin{aligned} Q_\theta(\theta|\overline{\lambda}) &= \sum_{i=1}^{N} Q_{\theta_i}(\theta_i|\overline{\lambda}) \\ &= \sum_{i=1}^{N} \left[\sum_{t=1}^{T} \sum_{k=1}^{K} c_{ikt} \log \omega_{ik} \mathcal{N}(x_t|\theta_{ik}) \right] \end{aligned} \tag{45}$$

where $\xi_{ijt} = \Pr(s_{t-1} = i, s_t = j|\mathbf{X}, \overline{\lambda})$ is the probability of making a transition from state i to state j at time t given that the model $\overline{\lambda}$ generates $\mathbf{X}$, and c_{ikt}, defined as

$$c_{ikt} = \gamma_{it} \frac{\overline{\omega}_{ik} \mathcal{N}(\mathbf{x}_t|\overline{m}_{ik}, \overline{r}_{ik})}{\sum_{l=1}^{K} \overline{\omega}_{il} \mathcal{N}(x_t|\overline{m}_{il}, \overline{r}_{il})} \tag{46}$$

is the probability of being in state i with the mixture component label k at time t given that the model $\overline{\lambda}$ generates $\mathbf{x}_t$, with $\gamma_{it} = \Pr(s_t = i|\mathbf{X}, \overline{\lambda})$. Both probabilities can be computed at each EM iteration using the forward–backward algorithm [15]. We can recognize from (46) that the membership function c_{ikt} has a similar form as was seen for c_{kt} in (35) for the mixture Gaussian case.

Similar to the mixture Gaussian case, estimating the mode of the posterior density requires the maximization of the auxiliary function, $R(\lambda|\overline{\lambda}) = Q(\lambda|\overline{\lambda}) + \log g(\lambda)$. The form chosen for the prior density $g(\lambda)$ in (42) permits independent maximization of each of the following $(2N + 1)$ HMM parameter sets: $\{\pi_1, \ldots, \pi_N\}$, $\{a_{i1}, \ldots, a_{iN}\}_{i=1,\ldots,N}$, and $\{\theta_i\}_{i=1,\ldots,N}$. The MAP auxiliary function

$$R(\lambda|\overline{\lambda}) = R_\pi(\pi|\overline{\lambda}) + \sum_i R_{a_i}(a_i|\overline{\lambda}) + \sum_i R_{\theta_i}(\theta_i|\overline{\lambda}) \tag{47}$$

where each term represents the MAP auxiliary function associated with the respectively indexed parameter sets. Equation (38) can be used to derive the reestimation formulas for π and $\mathbf{A}$ by applying the same derivations as were used for the mixture weights. The reestimation formulas (38)–(40) can also be used to maximize $R_{\theta_i}(\theta_i|\overline{\lambda})$ [56], [69].

So far we have only discussed MAP estimation for a single observation sequence. For multiple independent observation sequences, which is a more realistic situation in our applications, we can modify the auxiliary equation to include a summation over all data instances; the same reestimation equations can easily be extended [56].

2) Segmental MAP Estimation: By analogy with the segmental k-means algorithm [132], [86], a similar optimization criterion can be adopted. Instead of maximizing $g(\lambda|\mathbf{X})$, the joint posterior density of parameter $\lambda$and state sequence $\mathbf{s}$, $g(\lambda, \mathbf{s}|\mathbf{X})$, is maximized. The estimation procedure becomes

$$\begin{aligned} \hat{\lambda} &= \arg\max_{\lambda} \max_{\mathbf{s}} g(\lambda, \mathbf{s}|\mathbf{X}) \\ &= \arg\max_{\lambda} \max_{\mathbf{s}} f(\mathbf{X}, \mathbf{s}|\lambda) g(\lambda). \end{aligned} \tag{48}$$

$\hat{\lambda}$ is referred to as the *segmental MAP estimate* [102], [55], [56] of λ. Similar to the case for the segmental k-means algorithm, it is straightforward to prove that starting with any estimate $\lambda^{(m)}$, alternate maximization over $\mathbf{s}$ and λ gives a sequence of estimates with nondecreasing values of $g(\lambda, \mathbf{s}|\mathbf{X})$, i.e., $g(\lambda^{(m+1)}, \mathbf{s}^{(m+1)}|\mathbf{X}) \geq g(\lambda^{(m)}, \mathbf{s}^{(m)}|\mathbf{X})$ with

$$\mathbf{s}^{(m)} = \arg\max_{\mathbf{s}} f\left(\mathbf{X}, \mathbf{s} \,\middle|\, \lambda^{(m)}\right) \tag{49}$$

$$\lambda^{(m+1)} = \arg\max_{\lambda} f\left(\mathbf{X}, \mathbf{s}^{(m)} \,\middle|\, \lambda\right) g(\lambda). \tag{50}$$

The most likely state sequence $\mathbf{s}^{(m)}$ is decoded with the Viterbi algorithm. Maximization over λ can also be replaced by any *hill climbing* procedure over λ with the constraint, $f(\mathbf{X}, \mathbf{s}^{(m)}|\lambda^{(m+1)}) g(\lambda^{(m+1)}) \geq f(\mathbf{X}, \mathbf{s}^{(m)}|\lambda^{(m)}) g(\lambda^{(m)})$. The EM algorithm is once again a good candidate to perform this maximization using $\lambda^{(m)}$ as an initial estimate. It is straightforward to show that the forward–backward reestimation equations still holds if we set $\xi_{ijt} = \delta(s_{t-1}^{(m)} - i)\delta(s_t^{(m)} - j)$ and $\gamma_{it} = \delta(s_t^{(m)} - i)$, where $\delta(\cdot)$ denotes the Kronecker delta function.

D. Initial Prior Specification

In MAP-based HMM adaptation and other Bayesian learning scenarios, it critically depends on the choice of a prior pdf $g(\lambda|\varphi)$, which is often assumed to be a member of a preassigned family of prior distributions. In a strict Bayesian approach, the hyperparameter vector φ is also assumed known based on some subjective knowledge about λ. In reality, it is difficult to possess a complete knowledge of the prior distribution. An attractive compromise between the classical non-Bayesian approach, which uses no prior information and the strict Bayesian one is to adopt the empirical Bayes' (EB) approach [143], [117], [21]. By this we mean the hyperparameters are derived somehow from data. When replacing φ by any estimate derived from the already observed data, the previous and current data are linked in the form of a *two-stage sampling* scheme by a common prior pdf $g(\lambda|\varphi)$ of the unknown parameters λ.

Generally speaking, prior density estimation and the choice of density parameters depend on the particular application of interest. They also depend on the physical meaning of the variability or uncertainty we want to model and represent by using the prior pdf $g(\lambda|\varphi)$. For example, in speaker adaptation application, prior density $g(\lambda|\varphi)$ is used to model and represent the information of the variability of HMM parameters λ among a set of different speakers. In another application, for example, to build the context-dependent

models from context-independent models, the prior density $g(\lambda|\varphi)$ will be used to represent the variability of λ caused by different contexts. In the following, we highlight three approaches that we have used in the past several years and work quite well in a number of applications such as speaker adaptation, task adaptation, environment adaptation, etc.

If the training data set $\mathcal{X}$ for estimating hyperparameters φ is rich enough to cover the interested variability of speech signal, we can then divide $\mathcal{X}$ into different subsets $\mathcal{X} = \{\mathcal{X}^{(1)}, \mathcal{X}^{(2)}, \ldots, \mathcal{X}^{(Q)}\}$ according to the variability factors of interest. One can then estimate a set of HMMs, $\{\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_Q\}$ from the above sets of training data. One then pretend to view $\{\tilde{\lambda}_i\}$ as the random observations with the density $g(\lambda|\varphi)$. The *method of moment* detailed in [69], [72] can then be used to estimate φ. This provides a theoretically sound solution.

If the training data set $\mathcal{X}$ is not big enough, then we can use another method called *prior-weight initialization.* This method requires a set of seed models. Using these seed models and the conventional batch-mode ML training method, a set of statistics can be collected from a single pass through the training data $\mathcal{X}$. Using the collected statistics and a prior weight ϵ, the hyperparameters for CDHMMs can be specified as detailed in [71]. The prior weight ϵ controls the broadness of the prior pdf $p(\lambda|\varphi)$. One such example has been recently developed [76].

The third method is called τ*-initialization* [102], [55], [103]. For this method, with the assistance of a user-defined control parameter τ, the hyperparameter vector φ is specified directly from the parameters of existing seed models such that the mode of the derived prior pdf $p(\lambda|\varphi)$ is taken at the value of seed model parameters. Similar to the role of ϵ in prior-weight initialization, τ is used to control the broadness of $p(\lambda|\varphi)$. This method is attractive for those applications where only pretrained seed models are available and the training data $\mathcal{X}$ is not accessible during the prior specification.

V. Prior Evolution and On-Line Adaptation

The previously discussed MAP estimation methods imply batch algorithms that require processing the available data as a whole. In a variety of speech-recognition applications, it is desirable to process the data sequentially. The advantage of a sequential algorithm over a batch algorithm is not necessarily in the final result, but in computational efficiency, reduced storage requirements, and the fact that an outcome may be provided without having to wait for all the data to be processed. Moreover, the parameters of interest are sometimes subject to changes, e.g., they are time varying just like above-mentioned acoustic mismatch problem frequently encountered in real speech-recognition applications. In such cases, different data segments often correspond to different parameter values. Processing of all the available data jointly is no longer desirable, even if we can afford the computational load of the batch algorithm. To alleviate such problems, a sequential algorithm can be designed to adaptively track the varying parameters. This leads to an attractive adaptation scenario, which is known as the on-line (or *incremental*, *sequential*) adaptation. This scheme makes the recognition system capable of continuously adjusting to a new operational environment without the requirement of storing a large set of previously used training data. Among many possibilities (e.g., [110], [166], [70]–[73], [84], [41], [26], and [172]), Bayesian inference theory again provides a good vehicle to formulate and solve this problem. In this section, we will discuss one type of on-line adaptation approach, which is based on a key concept called *prior evolution.*

A. General Concept and Methodology

Suppose there are M speech units in a speech recognizer, each being modeled by a Gaussian mixture CDHMM. Consider a collection of such M CDHMMs $\Lambda = \{\lambda_q\colon\ q = 1, \ldots, M\}$, where $\lambda_q = (\pi^{(q)}, \mathbf{A}^{(q)}, \theta^{(q)})$ denotes the set of parameters of the qth N-state CDHMM used to characterize the qth speech unit. In a Bayesian framework, we intend to consider the uncertainty of the HMM parameters Λ by treating them as if they were random. Our prior knowledge about Λ is assumed to be summarized in a known joint *a priori* pdf $p(\Lambda|\varphi^{(0)})$ with *hyperparameters* $\varphi^{(0)}$, where $\Lambda \in \Omega$, Ω denotes an admissible region of the HMM parameter space. Such prior information may come from subject matter considerations. It can also be derived from previous experiences, e.g., training data $\mathcal{X}$, as we discussed in the previous section. Let $\mathcal{X}_1^n = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n\}$ be n independent sets of observation samples which are incrementally obtained and used to update our knowledge about Λ. Depending on different assumptions to make, constraints to apply, and knowledge sources to use, there are many ways to *evolve* $p(\Lambda)$. The central idea is that the intended evolving prior pdf $p_{\text{intend}}(\Lambda|\mathcal{X}_1^n)$ summarizes, in a way specified by each specific prior evolution scheme, the information inherited from the prior knowledge and learned from the observation data $\mathcal{X}_1^n$. From the evolving prior distribution, the intended inference and/or decision can be made. For example, we can derive a *point estimate* $\hat{\Lambda}$ from $p_{\text{intend}}(\Lambda|\mathcal{X}_1^n)$ (e.g., taking a mode) and then use the conventional plug-in MAP decision rule for recognition to achieve a better performance. This type of updating and use of Λ is known as on-line Bayesian adaptation in the speech community [70]–[73]. A block diagram of such an on-line adaptation scheme based on the concept of prior evolution is shown in Fig. 5.

Given a new block of input speech, feature extraction is first performed to derive the feature vector sequences used to characterize the speech input. It is followed by some kind of acoustic normalization to reduce the possible mismatch in the feature vector space. The processed feature vector sequences are then recognized based on the current set of HMMs. After the recognition of the current block of utterances, the prior pdfs for the relevant speech units, which are the results of the previous prior evolution step, are evolved to derive a set of *intended* posterior distributions, which will be served as the prior for the next round of prior evolution. By taking a point estimate from the evolved prior distributions, the related HMM parameters are adapted, and the updated models

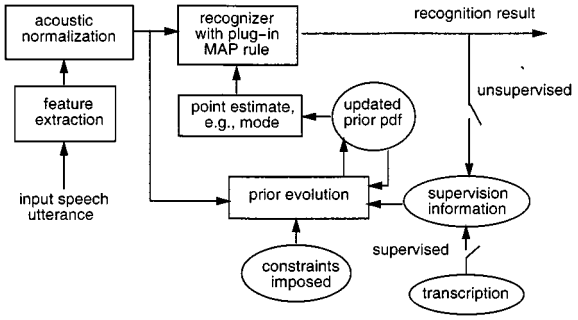

Fig. 5. On-line Bayesian adaptation based on prior evolution.

are used to recognize future input utterance(s). The prior evolution algorithm usually requires some form of supervision in terms of the word (or phone) transcription of the speech utterances. Such a transcription can be provided either by a human transcriber or by the correction made by the user on the recognized output during actual usage. This adaptation scheme is often called *supervised* adaptation. On the other hand, the supervision information can also be derived directly from the recognition results, and this is often referred to as *unsupervised* adaptation. For real-world applications, the unsupervised mode is usually more realistic and desirable.

In the following, we discuss the theoretical and practical issues related to several prior evolution schemes and explain how to use them for on-line adaptation of HMM parameters.

B. Prior Evolution Based on Recursive Bayesian Learning

One way to evolve $p(\Lambda)$ is to adopt the recursive Bayesian learning framework [159]

$$p(\Lambda|\mathcal{X}_1^n) = \frac{p(\mathcal{X}_n|\Lambda)\cdot p(\Lambda|\mathcal{X}_1^{n-1})}{\int_\Omega p(\mathcal{X}_n|\Lambda)\cdot p(\Lambda|\mathcal{X}_1^{n-1})\,d\Lambda}. \quad (51)$$

Starting the calculation of posterior pdf from $p(\Lambda|\varphi^{(0)})$, a repeated use of (51) produces a sequence of densities $p(\Lambda|\mathcal{X}_1^1)$, $p(\Lambda|\mathcal{X}_1^2)$, and so forth. It can be easily verified [159] that the above recursive way of computation for $p(\Lambda|\mathcal{X}_1^n)$ will give the same result as the one by using the following batch-mode computation:

$$p(\Lambda|\mathcal{X}_1^n) = \frac{p(\mathcal{X}_1^n|\Lambda)\cdot p\left(\Lambda|\varphi^{(0)}\right)}{\int_\Omega p(\mathcal{X}_1^n|\Lambda)\cdot p\left(\Lambda|\varphi^{(0)}\right)\,d\Lambda}. \quad (52)$$

If the computation in (51) can be carried out, this will give us an attractive sequential Bayesian estimation method.

Unfortunately, the implementation of this learning procedure for HMM raises some serious computational difficulties because of the nature of the *missing-data* problem caused by the underlying hidden processes, i.e., the state mixture component label sequence and the state sequence of the Markov chain for an HMM. It is well known that there exist no reproducing (natural conjugate) densities [159] for HMM. To illustrate this problem more clearly, let us consider the prior evolution for a single HMM λ. Let us begin with $g(\lambda)$ and consider what happens after a single adaptation utterance (sample) $\mathcal{X}_1 = \{\mathbf{X}\}$ is observed. For an observation sequence $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$, let $\mathbf{s} = (s_0, s_1, \ldots, s_T)$ be the unobserved state sequence and $\mathbf{l} = (l_1, l_2, \ldots, l_T)$ be the associated sequence of the unobserved mixture component labels. The posterior pdf of λ after observing $\mathbf{X}$ is

$$p(\lambda|\mathbf{X}) \propto \sum_{\mathbf{s},\mathbf{l}} \left\{ \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t}\omega_{s_t l_t}\mathcal{N}(\mathbf{x}_t|m_{s_t l_t}, r_{s_t l_t}) \right\} \cdot g(\lambda) \quad (53)$$

where the summations are taken over all possible state and mixture component label sequences. So the exact posterior pdf $p(\lambda|\mathbf{X})$ is a weighted sum of the prior pdf $g(\lambda)$, which includes $(N\cdot K)^T$ terms. Successive computation of (51) introduces an ever-expanding combination of the previous posterior pdfs and thus quickly leads to the combinatorial explosion of terms. As a result, formal recursive Bayes' learning procedures of this kind are not feasible and some approximations are needed in practice. A general approximation procedure is proposed in [71], which is to apply the Bayes' recursion of (51) incrementally, with one or more observation samples considered at a time. It is followed by a suitable approximation to the resulting posterior pdf so as to obtain recursive estimates of the hyperparameters of the approximate posterior pdf. This is typically accomplished by restricting the approximated pdf to be in the class of conjugate pdfs of the *complete-data* distributions. One such approach called quasi-Bayes' (QB) learning for HMM has been developed in [69], [71], and [72].

C. Prior Evolution Based on Quasi-Bayes' Learning

The quasi-Bayes' procedure is an approximate solution motivated by aiming at achieving computational simplicity while still maintaining the flavor of the formal Bayes' procedure. In the context of finite mixture distribution identification, the quasi-Bayes' approach was originally proposed by Makov and Smith [116], [157] to conduct recursive Bayes' estimation of the mixture coefficients while the mixture components are assumed fixed. In the sense that the approximate posterior distribution has a mean identical to that of the true posterior distribution, the convergence properties were established. We first adopted this approach to on-line adaptation of the mixture coefficients in the tied-mixture HMM case [70]. It was then extended to incremental adaptive learning of all of the CDHMM parameters in [71], [72]. In the following, we will explain our quasi-Bayes' learning framework in detail.

Depending on different assumptions to make and constraints to apply, we have studied several ways of defining $p(\Lambda|\varphi^{(0)})$. The simplest case is to assume λ_q's are independent, i.e.,

$$p\left(\Lambda\,\middle|\,\varphi^{(0)}\right) = g_1\left(\Lambda\,\middle|\,\varphi^{(0)}\right) = \prod_{q=1}^{M} g\left(\lambda_q\,\middle|\,\varphi_q^{(0)}\right) \quad (54)$$

where $g(\lambda_q|\varphi_q^{(0)})$ takes the same form as (42) , and $\varphi^{(0)} = \{\varphi_q^{(0)}\colon q = 1, 2, \ldots, M\}$. This class of prior distributions

$\{g_1(\cdot)\}$ actually constitutes a conjugate family of the *complete-data* density and is denoted as $\mathcal{P}_1$.

Under the above independence assumption, each model can only be adapted if the corresponding speech unit has been observed in the current adaptation data. Consequently, only after all units have been observed enough times can all of the HMM parameters be effectively adapted. To enhance the efficiency and the effectiveness of Bayes' adaptive learning, it is desirable to introduce some constraints on the HMM parameters. By this means, all the model parameters can be adjusted at the same time in a consistent and systematic way, even though some units are not seen in the adaptation data. One way to achieve the above objective is to explicitly consider the correlation of HMM parameters corresponding to different speech units (e.g., [99], [160], [180], [149], and [72]). For example, we can assume that the covariance matrices of HMMs, $\{\Sigma_{ik}^{(q)}\}$, are known. The initial prior pdf of Λ (excluding $\{\Sigma_{ik}^{(q)}\}$) is then assumed to be

$$p\Big(\Lambda\Big|\,\varphi^{(0)}\Big)=g_2\Big(\Lambda\Big|\,\varphi^{(0)}\Big)=g(\mathbf{m})\prod_{q=1}^{M}g(\lambda_q') \tag{55}$$

where

$$g(\lambda_q')\propto\prod_{i=1}^{N}\left\{\left[\pi_i^{(q)}\right]^{\eta_i^{(q)}-1}\cdot\left(\prod_{j=1}^{N}\left[a_{ij}^{(q)}\right]^{\eta_{ij}^{(q)}-1}\right)\cdot\left(\prod_{k=1}^{K}\left[\omega_{ik}^{(q)}\right]^{\nu_{ik}^{(q)}-1}\right)\right\} \tag{56}$$

is the product of a series of Dirichlet pdf (sometimes called multivariate beta pdf), and thus takes the special form of a matrix beta pdf [119] with sets of positive hyperparameters of $\{\eta_i^{(q)}\}$, $\{\eta_{ij}^{(q)}\}$, $\{\nu_{ik}^{(q)}\}$, and

$$g(\mathbf{m})=\mathcal{N}(\mathbf{m}|\boldsymbol{\mu},\,\mathbf{U}) \tag{57}$$

has a joint normal pdf with mean vector $\boldsymbol{\mu}=\text{vec}\{\mu_{ik}^{(q)}\}$ and covariance matrix $\mathbf{U}$ [99]. Here, we denote $\lambda_q'=(\pi_i^{(q)},\,a_{ij}^{(q)},\,\omega_{ik}^{(q)})$ and define $\mathbf{m}=\text{vec}\{m_{ik}^{(q)}\}$ to be the collection of the mean vectors of all the Gaussian mixture components of CDHMMs and denoted simply by an operator "vec." This class of prior distributions, $\{g_2(\cdot)\}$, constitutes another conjugate family of the *complete-data* density and is denoted as $\mathcal{P}_2$. In the following discussion, we will use $\mathcal{P}$ to refer to either $\mathcal{P}_1$ or $\mathcal{P}_2$, and its true meaning can easily be inferred from the context.

Consider at time instant n an adaptation set $\mathcal{X}_n=\{\mathbf{x}_n^{(q,r)}\}$ and prior knowledge about Λ approximated by $g(\Lambda|\varphi^{(n-1)})\in\mathcal{P}$. Here, $\mathbf{x}_n^{(q,r)}$ denotes the rth adaptation observation sequence of length $T_n^{(q,r)}$ associated with the qth speech unit, and each unit has $W_q^{(n)}$ such observation sequences. Let $\mathcal{Y}_n=(\mathcal{X}_n,\,\mathcal{Z}_n)$ denote the associated *complete-data* and $\mathcal{Z}_n=\{\mathbf{s}_n^{(q,r)},\,\mathbf{l}_n^{(q,r)}\}$ be corresponding *missing data*, where $\mathbf{s}_n^{(q,r)}$ denotes the unobserved state sequence and $\mathbf{l}_n^{(q,r)}$ is the sequence of the unobserved mixture component labels corresponding to the observation sequence $\mathbf{x}_n^{(q,r)}$. Given the set of observation sequences $\mathcal{X}_n$ and the above prior pdf $g(\Lambda|\varphi^{(n-1)})$, the QB procedure is, at each step of the recursive Bayes' learning, to approximate the true posterior distribution $p(\Lambda|\mathcal{X}_n)$

$$p(\Lambda|\mathcal{X}_n)=\frac{p(\mathcal{X}_n|\Lambda)\cdot g\left(\Lambda|\varphi^{(n-1)}\right)}{\displaystyle\int_\Omega p(\mathcal{X}_n|\Lambda)\cdot g\left(\Lambda|\varphi^{(n-1)}\right)\,d\Lambda} \tag{58}$$

by the "closest" tractable distribution $g(\Lambda|\varphi^{(n)})$ within the given class $\mathcal{P}$, under the criterion of both distributions having the same (local) mode [71], [72]. Here $\varphi^{(n)}$ denotes the updated hyperparameters after observing the samples $\mathcal{X}_n$. This idea is schematically illustrated in Fig. 6. More specifically, for any given estimate $\overline{\Lambda}$ let us define the auxiliary function

$$R(\Lambda|\overline{\Lambda})=\rho\cdot\log\,g\Big(\Lambda\Big|\,\varphi^{(n-1)}\Big)+E[\log\,p(\mathcal{Y}_n|\Lambda)|\mathcal{X}_n,\,\overline{\Lambda}] \tag{59}$$

where $0<\rho\le 1$ is a forgetting factor to be explained later and $\rho=1$ means that there is no forgetting. By choosing the initial prior pdf to be the conjugate family of the *complete-data* density, it can be verified that with an appropriate normalization factor C, such that

$$C\cdot\exp\{R(\Lambda|\overline{\Lambda}\}$$

belongs to the same distribution family as $g(\Lambda|\varphi^{(n-1)})$, and thus is denoted as $g(\Lambda|\hat{\varphi})$ with the hyperparameters $\hat{\varphi}$ detailed in [71] and [72]. By repeating the following EM steps, we can get a series of approximate pdf with the form $g(\Lambda|\hat{\varphi})$ whose mode is approaching the mode[2] of the true posterior pdf $p(\Lambda|\mathcal{X}_n)$.

E-step: Compute $R(\Lambda|\Lambda^{(n-1,l-1)})$ as in (59);

M-step: Choose

$$\Lambda^{(n-1,l)}=\arg\max_{\Lambda}R\Big(\Lambda\Big|\,\Lambda^{(n-1,l-1)}\Big) \tag{60}$$

where $l=1,\,2,\,\ldots,\,L$ is the iteration index, L is the total number of EM iterations performed, and $\Lambda^{(n-1,l)}$ is the estimated parameter at iteration l, with $\Lambda^{(n,0)}=\Lambda^{(n-1,L)}=\Lambda^{(n)}$ being the initial estimate at the beginning of the next EM iteration.

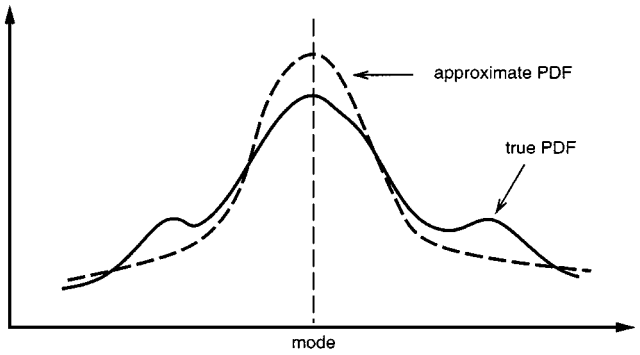

Fig. 6. Schematic illustration of quasi-Bayes' procedure.

[2]Strictly speaking, EM algorithm [37] can only guarantee the mode of the approximate pdf to approach a local maximum of the true posterior pdf in (58).

Thus, the hyperparameters $\varphi^{(n)}$ are obtained at the last (actually Lth) EM iteration to satisfy

$$g\left(\Lambda \,\middle|\, \varphi^{(n)}\right) \propto \exp\left\{R\left(\Lambda \,|\, \Lambda^{(n-1,\,L-1)}\right)\right\}. \qquad (61)$$

In this way, the old prior $g(\Lambda|\varphi^{(n-1)})$ is *evolved* to the new prior $g(\Lambda|\varphi^{(n)})$ via the adaptation data $\mathcal{X}_n$. The CDHMM parameters $\Lambda^{(n)} = \Lambda^{(n-1,\,L)}$ are then updated accordingly as in (60) by taking the mode of $g(\Lambda|\varphi^{(n)})$ as detailed in [71] and [72]. The updated HMMs are used in plug-in MAP rule to recognize future input utterance and this completes one step of on-line QB adaptation.

It is clear from the above discussion that using the concept of *density approximation* the QB algorithm is designed to incrementally update the hyperparameters on the approximate posterior distribution. Actually, the posterior distribution can further be manipulated, for example, if the initial prior knowledge is too strong or after a lot of adaptation data have been incrementally processed, the new adaptation data usually have only a small impact on parameter updating in incremental learning. To continuously track the variations of the model parameters corresponding to the new data, some *forgetting mechanisms* are needed to reduce the effect of past observations relative to the new input data. In the above prior evolution procedure, we actually introduced an *exponential forgetting* scheme by using a forgetting coefficient ρ as shown in (59). This is analogous to that proposed in [173] and [96].

The exponential forgetting is expected to be helpful for handling the slow changes of acoustic conditions between consecutive utterances by deemphasizing the contribution of the history data. For the abrupt (or fast) changes of condition, a *fast forgetting mechanism* is more helpful. Such a forgetting mechanism called *hyperparameter refreshing* was proposed in [71]. It can be roughly viewed as inflating the variance of $g(\Lambda|\varphi^{(n)})$ while maintaining the mode unchanged. It can also be viewed as an additional evolution step: from $g(\Lambda|\varphi^{(n)})$ to $g(\Lambda|\hat{\varphi}^{(n)})$ via the *variance inflation* or *hyperparameter refreshing*. Consequently, we obtain essentially a posterior distribution $p_{\text{intend}}(\Lambda|\mathcal{X}_1^n)$, which is different from the true posterior distribution $p_{\text{true}}(\Lambda|\mathcal{X}_1^n)$ but includes the information needed for adaptation from the observation data $\mathcal{X}_1^n$. The difference between the algorithms in [71] and [72] lies mainly in the fact that different constraints on HMM parameters are applied.

Recently, inspired by the above general QB framework and the general approximate recursive Bayesian learning framework in [20], a sequential learning method of mean vectors of CDHMM based on a finite mixture approximation of their prior/posterior densities has also been investigated [84]. More recently, by adopting a simple transformation (i.e., bias for mean vector and scaling for variance of CDHMM) and assuming a specific prior pdf for these transformation parameters, such a simple "transformation-based" QB adaptation algorithm has been developed in [26] by using the general QB framework in [71]. This algorithm can be viewed as another way of prior evolution with the above-mentioned linear constraints imposed. As a final remark, the above QB framework is also flexible enough to include the batch-mode MAP estimation as a special case, which can be viewed as a one-step prior evolution with QB, followed by a point estimate (taking a mode) from the evolved prior.

D. Multiple-Stream Prior Evolution and Posterior Pooling

In addition to the above method of prior evolution, we can also, for example, assume Λ to evolve in a more *constrained* way as $\Lambda^{(n)} = H_n(\Lambda^{(0)})$, where H_n represents a mapping from $\Lambda^{(0)}$ to $\Lambda^{(n)}$ and can be *incrementally* learned from the observation data $\mathcal{X}_1^n$. Then from $p(\Lambda|\varphi^{(0)})$, we can derive a new posterior distribution $p_{\text{intend}}(\Lambda|\mathcal{X}_1^n) = p(\Lambda^{(n)})$ as the result of prior evolution. Of course, there are other ways to evolve $p(\Lambda)$. Each leads to a different on-line adaptive learning algorithm. Moreover, the prior evolution can start from either a single prior pdf, or more generally, different prior pdfs for different schemes. Depending on the specific meaning of the prior pdf and the way of prior evolution, different schemes might reflect different aspects of the learning. A natural way of obtaining an enhanced learning algorithm is to simultaneously maintain multiple streams of prior evolution. During the process of the prior evolution, we can design a *posterior pooling* scheme, which combines different streams of evolved pdfs to derive an intended pdf for further inference or decision-making. Such a framework called *multiple-stream prior evolution and posterior pooling* has been recently developed in [73] and is briefly described in the following.

The use of multiple stream prior evolution is well motivated. This is because the speech signal is very rich, which includes the desirable linguistic information for recognition as well as many other undesirable variation. A *multifacet learning* algorithm can thus be designed to elicit from a rich set of training data $\mathcal{X}$ a set of prior distributions, $\{p^{(i)}(\Lambda|\varphi_i^{(0)}),\, i = 1,\, 2,\, \ldots,\, I\}$. Each $p^{(i)}(\Lambda|\varphi_i^{(0)})$ reflects how HMM parameters Λ varies according to one type of variability factors (e.g., speakers, speaking styles, data capturing and transmission conditions, etc.). We can treat each $p^{(i)}(\Lambda|\varphi_i^{(0)})$ as a *knowledge source*, which reflects one aspect of the speech signal.

After we have prepared the set of $\{p^{(i)}(\Lambda|\varphi_i^{(0)})\}$ from training data $\mathcal{X}$, we can then use them to *compose* and *derive* a condition-dependent distribution $p_{\text{intend}}(\Lambda|\mathcal{X}_{\text{new}})$ guided by task specifications and a small amount of *condition-dependent* adaptation data (possibly derived from test data) $\mathcal{X}_{\text{new}}$. To achieve this goal, we can first compose an intended distribution

$$\tilde{p}_{\text{intend}}(\Lambda|\mathcal{X}_{\text{new}}) = \sum_i \epsilon_i \times p^{(i)}\left(\Lambda \,\middle|\, \varphi_i^{(0)}\right) \qquad (62)$$

where ϵ_i ($0 \le \epsilon_i \le 1$ and $\sum_i \epsilon_i = 1$) is the *fusion weight* to control the *relative importance* of the different knowledge sources $p^{(i)}(\Lambda|\varphi_i^{(0)})$. The ϵ_i's can either be automatically trained from the adaptation data $\mathcal{X}_{\text{new}}$ or just be specified according to task specifications and modeling intention. Then

we can use a manageable distribution $p(\Lambda|\hat{\varphi})$ to approximate $\tilde{p}_{\text{intend}}(\Lambda|\mathcal{X}_{\text{new}})$ by minimizing the *Kullback–Leibler directed divergence* [97] as follows:

$$\hat{\varphi} = \arg\min_{\varphi} \int \tilde{p}_{\text{intend}}(\Lambda|\mathcal{X}_{\text{new}}) \log \frac{\tilde{p}_{\text{intend}}(\Lambda|\mathcal{X}_{\text{new}})}{p(\Lambda|\varphi)} \, d\Lambda. \tag{63}$$

With $p(\Lambda|\hat{\varphi})$, we can derive a point estimate (e.g., taking a mode) of Λ and then to use the *plug-in MAP decision rule* to construct a speech recognizer. In [73], we have shown an implementation of the above *information fusion* method when we only consider the uncertainty of the mean vectors of CDHMMs.

Alternatively, we can first evolve $p^{(i)}(\Lambda|\varphi_i^{(0)})$ by using the adaptation data $\mathcal{X}_{\text{new}}$ and an appropriate prior evolution method to obtain a set of intended distributions $\{p^{(i)}_{\text{intend}}(\Lambda|\mathcal{X}_{\text{new}})\}$. Then the above information fusion technique can be used to derive $p(\Lambda|\hat{\varphi})$ and to construct the speech recognizer accordingly. If the application involves many utterances during the real use of the ASR system, the above scheme can be operated in an incremental mode. This technique of *multiple-stream prior evolution and posterior pooling* can thus be used to continuously improve the ASR performance with the increasing amount of condition-dependent speech data. In [73], we have proposed several architectures for *multiple-stream prior evolution*. In a case study where two-stream prior evolution is used and only the uncertainty of the mean vectors of the CDHMMs are considered, good results are obtained for efficient speaker adaptation application.

The above approach of prior evolution and posterior pooling opens up many new research opportunities. The key to the success of these approaches depends on whether the imposed constraints really exist in the entities under investigation. By using multiple-stream framework, we can always exploit multiple sources of knowledge and/or apply different kinds of constraints to facilitate learning. It is believed that the best setup will depend on the purpose of modeling and learning as well as the nature of the specific applications. Intelligent use of the previously discussed flexible tools for different purposes in different applications will be an important part of the future research.

VI. ML/MAP Estimation of Structural Parameters

The Bayesian adaptation approach we discussed so far provides an optimal mathematical framework for combining information in a general set of stochastic models and a specific set of adaptation data. However, in order to improve efficiency and effectiveness when dealing with adaptation of many parameters with a very limited set of data, new techniques have been proposed recently in exploring structures embedded in the feature and model spaces. Successful usage of these structures allows us to incorporate this newly available set of structural parameters into many of the parameter estimation and adaptation algorithms. In the following, we discuss a number of such techniques, including model parameter transformation, interpolation, correlation, and normalization. We will also briefly discuss a way to align prior information in a hierarchical tree of the models for *structural Bayesian adaptation*. The combination of structure and model parameter estimation opens up brand new ways to design plug-in MAP decision rules. Some of them are shown in the left half of the roadmap in Fig. 3. This new set of auxiliary structures also provides a mathematical framework to approximate some of the missing channel information, such as speakers and microphones, illustrated in Fig. 1 but ignored in the simplified source-channel model shown in Fig. 2.

A. Model Transformation and Interpolation

The most studied structures are those defined through constraints on the model parameters. Such methods bind the models in ways that all the parameters are adjusted simultaneously according to the predetermined set of constraints, e.g., *multiple regression analysis* as suggested in the classical paper by Furui [49]. Instead of local estimation or adaptation for HMM parameters, the transformation-based approaches capture some global behavior of the parameter space. Therefore, it works better for small-size adaptation data or no data at all in the case of unsupervised adaptation (or compensation). This rich family of techniques are highlighted in the left half of the roadmap in Fig. 3. In the upper part, ML estimation and interpolation are illustrated. A MAP version of the chart is summarized in the bottom left part and a hybrid version, in which ML estimation and MAP adaptation of HMM parameters are combined, is shown in the bottom right.

In general, the above constraint set is introduced through some form of parameter transformation, $\Lambda_{\overline{X}} = F_{\Phi}(\Lambda_X)$, in which Λ_X and $\Lambda_{\overline{X}}$ denote the original and the transformed parameter vectors, respectively, $F_{\Phi}(\cdot)$ is a transformation of interest, and Φ is a small set of transformation parameters characterizing the model transformation. Then given a set of training/adaptation data $\mathcal{X}$, we obtain an ML estimate of Φ and hence $\Lambda_{\overline{X}}$ by solving

$$\hat{\Phi} = \arg\max_{\Phi} p(\mathcal{X}|\Lambda_X, \Phi) \tag{64}$$

where $p(\mathcal{X}|\Lambda_X, \Phi) = p(\mathcal{X}|\Lambda_{\overline{X}})$ is the observation pdf with the transformed parameter $\Lambda_{\overline{X}}$. Alternatively, one can assume the transformation parameter Φ to be a random variable and specify a prior density $p(\Phi)$ to capture some prior knowledge about Φ. Then the MAP estimate of Φ can be solved as

$$\hat{\Phi} = \arg\max_{\Phi} p(\mathcal{X}|\Lambda_X, \Phi) \cdot p(\Phi). \tag{65}$$

One popular way is to directly impose a regression constraint on the model mean vectors and estimate the linear regression parameters, as in MLLR, using an EM algorithm (e.g., [109]). Just like MAP, MLLR has been adopted by many recent ASR systems for its simplicity and effectiveness. The readers are referred to a recent review [176] covering this rich family of techniques. In the MLLR framework, variance can also be estimated similarly [53]. Constrained estimation of Gaussian mixture parameters can be considered as a constrained MLLR approach and has been studied in [39]. When more adaptation data are available, more transfor-

mations are needed. These transformations are made class specific so that different units can be adapted differently depending on their corresponding acoustic or linguistic classes (e.g., [110]). As we mentioned above, MAP, instead of ML, can be used for estimating the regression parameters as shown in [154]. To obtain a closed-form MAP solution, a family of *elliptically symmetricmatrix variate priors* was adopted [30], [63] to specify the prior of the parameters in the linear regression matrix. The above MLLR and MAPLR approaches work well, especially in the case of adaptation with a small amount of data as well as in unsupervised adaptation (e.g., [175] and [154]).

Another way to accomplish speaker adaptation is through a simple affine transformation between reference and adaptive speaker feature vectors. It is then translated into a bias vector and a scale matrix, which can be estimated with an EM algorithm in the adaptation process [39], [147], [182], [144]. Model-based bias adaptation has been a well-studied topic for robust speech recognition. Some applications to compensation will be discussed in the next section.

As expected, when combined with Bayesian adaptation, simple transformations show a good adaptation efficiency (for short adaptation data) and a good asymptotic property (converging to speaker-dependent models). When modeling channel as a bias transformation (e.g., [147]), MAP adaptation can also be used to improve recognition performance (e.g., [23] and [169]). Stochastic matching has been combined with MAP adaptation in [24].

Yet, another alternative to accomplish the above is to define a *vector field* for the set of mean parameters and assume that the adaptation data vectors are used to transfer the vector field of the reference model to that of the new speaker in a consistent manner so that the mean vectors of unseen units can be interpolated from the estimated mean vectors of observed units. It can be considered as a constraint to preserve the structure of the vector fields before and after the vector transfer and smoothing operation. For the mean vector of each unseen unit model, such an interpolation is usually confined to a neighborhood in the vector field of the reference model so as to improve the robustness of the transfer [65], [128]. When combined with Bayesian adaptation, *vector field smoothing* (VFS) has been shown effective for both batch [168] and incremental adaptation [166]. SM and VFS have also been compared and combined with MAP adaptation to improve speech recognition, as shown in [25].

Before we close this important subject, it is noted that formal joint MAP estimation of the transformation and HMM parameters can also be obtained via

$$(\hat{\Lambda}_X, \hat{\Phi}) = \arg \max_{(\Lambda_X, \Phi)} p(\mathcal{X}|\Lambda_X, \Phi) \cdot p(\Lambda_X, \Phi) \qquad (66)$$

with $p(\Lambda_X, \Phi)$ being the joint prior of the two parameter sets. Care is needed when solving through an iterative EM procedure [155] that finds $\hat{\Phi}$ given Λ_X, then $\hat{\Lambda}_X$ given $\hat{\Phi}$. It was found that the results obtained with the joint estimation procedure are better than those obtained with MAPLR of Φ or with MAP of Λ_X alone in all adaptation sizes tested [155]. It also retains the nice asymptotic properties of MAP estimation. Other possibilities of combining existing techniques, such as on-line recursive Bayesian learning of tree-structured transformation [172], have been recently studied. More are expected as implied in the bottom block of Fig. 3.

B. Parameter Correlation

In a conventional HMM-based Bayesian adaptation framework, the parameters between different HMMs are usually assumed independent. Therefore, each HMM can only be adapted if the corresponding unit has been observed in the adaptation data. Since it is unlikely to have observed all the units enough times in a small adaptation set, only a small number of parameters can be effectively adapted. It is, therefore, desirable to introduce some parameter correlation or tying so that all the model parameters can be adjusted at the same time in a consistent manner, even though some units are not seen in adaptation data.

Because defining a joint distribution, having an HMM marginal for each unit, is a difficult job, two alternatives have been used. First, instead of assigning a single label to each adaptation data segment, multiple labels can be used. For example, a speech segment can be associated with both context-dependent and context-independent labels so that it can be used to adapt both types of HMMs (Type II training in [55]). Second, a correlation structure between parameters can be established and the correlation parameters can be estimated when training the general models or constructing tying structures [167]. Parameters of unseen units can then be adapted accordingly (e.g., [99], [32], [180], [167], [149], [22], and [72]). *Regression-based model prediction* (RMP) [49], [31] combined with a Bayesian approach has also been studied [2]. This area of work is summarized in the bottom right block of the technology roadmap shown in Fig. 3. Another way is to introduce correlation through a hierarchical structure as in the extended MAP [180] and the structural MAP [151], [26] approaches. The MAP algorithm is in spirit similar to an ML version, the so-called *autonomous model complexity control* [150], to determine the complexity of the models based on the size of the training data.

C. Structural Bayesian Adaptation

The definition of a structure to aid MAP estimation is a key procedure in the structural Bayes' approach [151]. Consider for the set G of all the Gaussian mixture components in a set of CDHMMs a *tree structure* where $K$is the total number of layers or the depth of the tree. Each node in the Kth layer (leaf node) corresponds to one Gaussian mixture component. The root node (the first layer) corresponds the whole set G of the mixture components. Each intermediate node corresponds to a subset of G, and each of its subordinate leaf nodes corresponds to an element of a subset.

By assuming that the prior knowledge in a tree node can be used to construct prior density needed for MAP estimation of all the parameters in the successive child nodes, a new *structural maximum a posteriori* (SMAP) algorithm [151], [152] has been developed for speaker and environment adaptation.

It allows simultaneous adaptation of all the mixture Gaussian parameters, even with only one adaptation utterance.

Three key steps are required in formulating the proposed SMAP approach. First, a tree is used to characterize the acoustic space represented by the HMM parameters. In [151], an information theoretical criterion is used to cluster all the Gaussian mixture component densities typically used to model state observation densities in HMM. Next, given all the density clusters used to characterize nodes in a tree, we need to find a Gaussian density to summarize all the Gaussian components in the cluster so that the likelihood of a sequence of observation vectors representing the adaptation data can be evaluated at the node level and, therefore, the MAP estimate at any node in the tree can be computed. For the third step, the prior density at each tree node needs to be defined. In order to use every observation sample to estimate all the HMM parameters, we use a *hierarchical prior evolution approximation* by assuming that the hyperparameters characterizing the prior density at each node are evaluated based on the knowledge embedded in the prior density of its parent node. Once the three key steps are established, the SMAP estimation algorithm is then derived.

The SMAP procedure was shown to be effective for supervised batch adaptation [151], unsupervised incremental adaptation [152], and combined supervised rapid adaptation and unsupervised incremental adaptation [152] in order to reduce the amount of adaptation data needed to achieve a reasonable level of performance and to improve performance degradation in mismatch conditions. Some recent work in extending SMAP to handle structural parameters such as MLLR, called SMAPLR [156], has also demonstrated the effectiveness of structural Bayesian adaptation approaches.

D. Adaptation and Normalization

Structural parameters, such as those in affine transformations, can also be used for normalizing the influence of speakers, channels, and environments as suggested in Fig. 5 so that the heterogeneity embedded in a large set of acoustic training data can be reduced and a compact set of acoustic models can be estimated. The normalization process can be carried out both in the feature, and model spaces (e.g., [181]). For example, the popular *cepstral mean normalization* (CMN) algorithm [6] can be applied to every training utterance to reduce some channel and speaker effect. *Codeword-dependent cepstralnormalization* (CDCN) [1] and its variations [114] can be considered as extensions of CMN. Speaker normalization through *vocal tract length normalization* (VTLN) using *frequency warping* has also been proposed (e.g., [108]). ML-based feature normalization, such as *signal bias removal* (SBR) [135] and *stochastic matching* (SM) [147], which was originally developed for compensation, can also be performed before model training or adaptation (e.g., [181]). A frame-synchronous stochastic matching algorithm has also been proposed for real-time processing [35]. Both piecewise linear transformation (the so-called *metamorphicnormalization* [17]) and linear regression transformation [130] have been applied to map training data of certain selected training speakers based on a small amount of adaptation data to obtain an artificial bigger size training set for speaker adaptation. Instead of mapping training data, linear regression transformations have also been used to map the selected speaker cluster models based on small-size adaptation data and then compose a speaker-adaptive model [54]. Furthermore, these transformations have been used to normalize the so-called *irrelevant variabilities* in an integrated ML *speaker adaptive training* (SAT) scheme [5] to obtain a set of generic speech models. Since the variations contributed by speakers is reduced first, the resulted speaker independent models are in principle more compact, i.e., requiring less parameters. However, in this case, an MLLR or MAP speaker adaptation routine usually has to be performed to achieve a good performance for a new speaker. Such procedures can also be carried out on speaker clusters, such as female and male groups, and on other channel factor classes shown in Fig. 1. It is important to know the interactions among normalization, compensation, and adaptation procedures to maximize the utility for designing plug-in decision rules for ASR.

VII. Adaptation, Compensation, and Robustness

We have discussed a number of parameter adaptation techniques to estimate the acoustic and language models for designing a plug-in MAP decoder for ASR. However, in many situations, such adaptive decision rules are still not capable of coping with the changing conditions and, therefore, mismatch from training to testing. The most effective way to handle mismatch seems to be finding invariant features so as to minimize the effect of acoustic mismatch between training and testing environments. Even though some features have been shown less affected by a certain type of distortion, such as linear microphone or channel effect, no feature has been discovered that is invariant across all adverse acoustic conditions. To circumvent this difficulty, a straightforward solution is to collect additional adaptation data in a specific testing condition and then to adapt the recognizer parameters accordingly to work in the prescribed scenario. A more realistic approach is to again perform adaptive learning during testing assuming no knowledge about the new acoustic conditions or the actual sentence (or transcription) spoken. This process is often referred to as compensation, as opposed to adaptation. Compensation can be considered as a form of unsupervised adaptation in which only the testing utterances are used. Many other names have also been adopted, e.g., *self adaptation*, *auto adaptation*, *instantaneous adaptation*, or stochastic matching (e.g., [180], [182], [147], [53], [164], and [123]). For robust speech recognition, compensation can be accomplished in the signal, feature and model spaces in order to reduce the distortions shown in Fig. 7 (adopted from [147]). The readers are referred to a recent review on the topic of feature and model compensation (e.g., [106]). In this section, we focus our discussion on the relationship between adaptation and compensation and on how to apply some of the adaptation techniques discussed in previous sections to improve robustness of speech recognizers. It is noted that algorithms originally developed for adaptation, such as

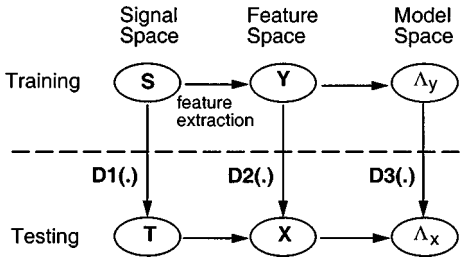

Fig. 7. Conceptual diagram of mismatch between training and testing.

MLLR, can also be applied directly to compensation. Similarly, techniques designed to handle compensation, such as model-based stochastic matching, can also be used for adaptation and normalization. For more general discussions on robust speech recognition, the readers are referred to a few recent publications (e.g., [90] and [158]).

One of the earliest studies on feature compensation is cepstral mean normalization [6], which removes the cepstral mean of each utterance before training and testing. CMN was shown to be robust to microphone and channel distortion in many systems. By making CMN more effective for different sounds in different speaking conditions, CDCN and its derived techniques [1], [114] were then developed. A simplified version known as signal bias removal or *signal conditioning* was shown to be effective for several applications (e.g., [182] and [135]). *Hierarchical spectral clustering* in designing vector quantization codebook for normalization has also been proposed [50]. Typically, a codebook is used to represent the reference acoustic space and then a set of biases can be derived to compensate cepstral difference between testing feature vectors and reference codebook. When no training data are available to create the codebook, a natural extension is to use the information embedded in the acoustic HMMs to aid the feature compensation process (e.g., [182] and [147]). In stochastic matching [147], which is in essence a model-based equalization algorithm, the entire set of HMMs is used to perform feature compensation and solved for the recognized sentence.

A. Self-Adaptation or Compensation

In the previous sections, we have discussed techniques that require a set of data to perform rapid batch or incremental adaptation. However, adaptation can also be performed at runtime on the testing data in an unsupervised manner. This process is often referred to as *self-adaptation.* The idea is to introduce additional parameters or structures to account for some models of mismatch in testing, and such parameters are to be estimated along with the recognized sentence during actual testing. This is an important way to enhance robustness toward varying environments, microphones, channels, and speakers. One way to do self-adaptation is through the Type III Bayesian adaptation [55] where the same testing utterance is used to obtain MAP estimator of all unit parameters without using the recognized unit labels. A second approach is through stochastic matching in which *nuisance parameters* for distortion, and the recognized sentences are estimated with an EM algorithm [147], [182]. The extended MAP approach [180] and MLLR/MAPLR have also been shown applicable to self-adaptation (e.g., [175] and [154]).

Because the amount of adaptation data is limited to the testing utterance itself in self-adaptation, constraints are needed to reduce the number of parameters to be adapted. For example, in stochastic matching, such a constraint is introduced through some form of parameter transformation $\Lambda_{\overline{X}} = F_\Phi(\Lambda_X)$, in which Λ_X and $\Lambda_{\overline{X}}$ denote the original and the transformed parameter vectors, respectively, $F_\Phi(\cdot)$ is a transformation of interest and Φ is a small set of transformation parameters characterizing the model transformation. Then self-adaptation amounts to solving the following optimization problem:

$$(\hat{W}, \hat{\Phi}) = \arg \max_{(W, \Phi)} p(X|W, \Phi, \Lambda_X) P(W) \qquad (67)$$

where the nuisance parameter Φ is solved together with performing plug-in MAP decoding by iterative algorithms [147], [53], [164].

B. Model Compensation

Although model-based feature compensation is effective in some situations, there are many types of distortion that cannot easily be realized by a simple feature transformation. Sometimes the exact distribution of the transformed feature vectors can not be derived in a useful form for decoding, i.e., a numerical procedure might be required. Model compensation provides an attractive alternative. For example, if the feature bias is time varying, i.e., $x_t = y_t + b_t$ with b_t being a *stochastic bias* [147], then the feature compensation vector can not be computed exactly. If b_t is a random vector with mean vector μ_b and covariance matrix Σ_b and is independent from the speech features y_t, then it is equivalent to solving the bias density parameters by the following model transformations: $\mu_x = \mu_y + \mu_b$ and $\Sigma_x = \Sigma_y + \Sigma_b$. The nuisance parameters μ_b and Σ_b are solved either with a ML-based EM [147] or by a MAP-based EM algorithm [24]. Other structures can also be employed to reduce the number of parameters while improving compensation efficiency and effectiveness (e.g., [106]).

Other transformations, such as *hierarchical VQ* [50], *probabilistic spectrum fitting* [31], *context modulation* [181], *spectral equalization* [165], *affine transformation* [39], and the widely used MLLR [109], although originally developed for speaker adaptation, can also be used for model compensation. Care is needed when applying such adaptation algorithms for compensation. First, compensation is equivalent to adaptation without transcription for supervision, as shown in Fig. 5. Therefore, only reliable unsupervised adaptation algorithms can directly be used for compensation. Another concern is that the number of the parameters to be compensated should be limited because typically, only a small amount of testing utterances is used for compensation. Tying of transformation parameters (e.g., [40] and [110]) needs to be considered in order to have an effective compensation.

If the additive bias is an HMM, then x_t is also an HMM. However, the topology and the state observation densities of

x_t can be very different from those of y_t, the original speech HMM. Therefore, new models of x_t need to be estimated and implemented during decoding. This family of algorithms is known as *model decomposition* [170], *parallel model combination* (PMC) [52], or *model composition* [118]. The original algorithms [170], [53] were designed to handle additive noise only. It was later extended to cope with both additive and convolutional noises [51]. Other algorithms that estimate the joint densities of speech and noise have also been suggested (e.g., [45] and [144]). We believe compensation together with adaptation and normalization are some keys to future technology advances in robust speech recognition.

VIII. Robust Decision Rules

As we discussed in Section II, the plug-in MAP decoder minimizes the recognition error only if the form of the distributions of the data to be recognized and the corresponding parameters are known exactly. The above adaptation and compensation strategies improve the robustness of speech recognition systems by making the distribution $p_{\hat{\Lambda}}(\mathbf{X}|W)$ reflect more faithfully the true distribution of $p(\mathbf{X}|W)$ for utterance $\mathbf{X}$ to be recognized, while keeping the plug-in MAP classification and decision rules intact. Another possibility to improve the robustness of an ASR system is to modify the plug-in MAP decoder. This area has not attracted much research attention, partly because the dynamic programming-based search strategies for implementing the plug-in MAP decoder are by far the most efficient implementation for solving speech recognition solutions. Any modification of the prevailing DP search algorithm requires a considerable amount of work. However, there exist robust decision rules that can be implemented without changing too much of the existing DP-based algorithms.

Intuitively speaking, a decision strategy (rule) is called robust if it is not very sensitive to the previously discussed prior uncertainty (or distortions). The readers are referred to [95] for a formal definition of *decision rule robustness*. In the following two sections, we show two examples of such robust decision rules, namely, *minimax classification rule* and *Bayesian predictive classification* (BPC) *rule*, respectively. Both of them assume the following:

1) the distributions $p(\mathbf{X}|W)$ and $P(W)$ are known up to some specifiable parameters in the forms of $p_\Lambda(\mathbf{X}|W)$ and $P_\Gamma(W)$;
2) the *true* parameters of these distributions, Λ, and Γ, lie in a neighborhood of the estimated (or hypothetical) ones;
3) the, therefore, *prior uncertainty* can be modeled by defining an *uncertainty neighborhood* of the model parameters Λ and Γ and/or possibly a distribution of model parameters $p(\Lambda, \Gamma)$ on this *uncertainty neighborhood*.

With these assumptions, the specific minimax decision rule and predictive decision rule can be constructed accordingly to satisfy some desired robustness properties. To simplify our discussion, we further assume that we do not consider the uncertainty of $P(W)$ and use $P_{\Gamma_0}(W)$ as the language model, with Γ_0 being the set of language model parameters estimated from the training text data.

A. Minimax Classification

Let $\eta_\epsilon(\Lambda_0)$ denote the uncertainty neighborhood of the true model parameters Λ, i.e., $\Lambda \in \eta_\epsilon(\Lambda_0)$, where Λ_0 is the set of model parameters estimated from the training data $\mathcal{X}$ and ϵ can be viewed as a generic parameter to characterize the degree of the distortion. Then, we have

$$\mathcal{M}_\epsilon^* = \{p_\Lambda(\mathbf{X}|W) | \Lambda \in \eta_\epsilon(\Lambda_0)\} \tag{68}$$

where $\mathcal{M}_\epsilon^*$ is the set of distorted models. With $\mathcal{M}_\epsilon^*$, a functional, namely, an *upper bound* of the *worst case probability of classification error,* can be defined [122]. A decision rule that minimizes this functional is as follows:

$$\hat{W} = \arg\max_W \left[P_{\Gamma_0}(W) \cdot \max_{\Lambda \in \eta_\epsilon(\Lambda_0)} p_\Lambda(\mathbf{X}|W) \right]. \tag{69}$$

This is the so-called minimax classification rule, which was first studied by Merhav and Lee [122]. It can be solved in two steps. First, we estimate the underlying parameters using the ML approach within each neighborhood $\eta_\epsilon(\Lambda_0^{(W)})$, i.e.,

$$\hat{\Lambda}_W = \arg \max_{\Lambda \in \eta_\epsilon\left(\Lambda_0^{(W)}\right)} p_\Lambda(\mathbf{X}|W) \tag{70}$$

where $\Lambda_0^{(W)}$ denotes pretrained model parameters for word W. Then we apply the plug-in MAP decision rule, with $\hat{\Lambda}_W$ replacing the original $\Lambda_0^{(W)}$. Therefore, conceptually, the minimax decision rule described in (69) can be viewed as a procedure that modifies the (plug-in) MAP decoder shown in (2) with an extra step as in (70) to find a modified point estimate in the neighborhood $\eta_\epsilon(\Lambda_0) = \{\eta_\epsilon(\Lambda_0^{(W)})\}$ of the original classifier parameters $\Lambda_0 = \{\Lambda_0^{(W)}\}$.

The above robust minimax classification rule makes no assumption about the form of the distortion. However, its efficacy does depend on an appropriate specification of the parameter uncertainty neighborhood $\eta_\epsilon(\Lambda_0) = \{\eta_\epsilon(\Lambda_0^{(W)})\}$. In the past several years, some other specific techniques have also been developed to implement the above *minimax decision rule* in HMM-based ASR systems (e.g., [124] and [82]). They are shown to be effective in dealing with noisy speech recognition and the mismatch caused by different recording conditions.

There are also other possibilities to model the admissible distortions $\mathcal{M}_\epsilon^*$. For example, if we use

$$\mathcal{M}_\epsilon^* = \{p_\Lambda(\mathbf{X}|W) | \Lambda = \mathcal{T}_\vartheta(\Lambda_0)\} \tag{71}$$

where $\mathcal{T}_\vartheta(\Lambda_0)$ denotes a specific transformation of Λ_0 with parameters ϑ. In this way, the uncertainty of Λ can be characterized by the uncertainty of ϑ. Then the *minimax decision rule* with respect to the above $\mathcal{M}_\epsilon^*$ will be

$$\hat{W} = \arg\max_W \left[P_{\Gamma_0}(W) \max_\vartheta p(\mathbf{X}|W, \Lambda = \mathcal{T}_\vartheta(\Lambda_0)) \right]. \tag{72}$$

The so-called *model-space stochastic matching* method described in [147] and [164] can be theoretically justified in this way.

B. Bayesian Predictive Classification

As we discussed before, minimax classification tries to handle the worst case mismatch by assuming a uniform distribution in the uncertainty neighborhood for all possible deviation from the nominal parameters Λ_0. Instead of assigning another *point* estimate $\hat{\Lambda}$ as done in the minimax classification rule discussed above, one can also *average out* the effect of the possible modeling and estimation errors by assuming a general prior pdf for Λ to characterize the parameter variability while making classification decisions. In this way, a new robust decision strategy can be derived and is often referred to as a Bayesian predictive classification rule (e.g., [126] and [74]).

The principle behind the BPC approach is quite straightforward. Because we assume no knowledge about the possible distortions, we thus rely on a quite general prior pdf to characterize the variability of the HMM parameters caused by the possible mismatches and errors in modeling and estimation. Let us consider the uncertainty of the model parameters Λ by treating them as if they were random. Our *prior uncertainty* about Λ is then assumed to be summarized in a known joint *a priori* density $p(\Lambda|\varphi)$, with $\Lambda \in \Omega_\Lambda$, where Ω_Λ denotes an admissible region of Λ and φ is the set of parameters of the prior pdf. In this way, we are essentially considering the following admissible distorted set of data model $\mathcal{M}_\epsilon^*$

$$\mathcal{M}_\epsilon^* = \{p_\Lambda(\mathbf{X}|W)|\Lambda \sim p(\Lambda|\varphi);\ \Lambda \in \Omega_\Lambda\} \tag{73}$$

where we can view ϵ as a parameter to characterize the broadness of the distribution $p(\Lambda|\varphi)$ or, equivalently, the degree of the distortion. If we want to account for model parameters' uncertainty in *recognition*, an *optimal Bayes' solution* , namely, BPC, exists. It selects a speech recognizer to minimize the *overall recognition error* (this is when the average is taken both with respect to the sampling variation in the expected testing data and the uncertainty described by the prior distribution). Such a BPC rule operates as follows:

$$\begin{aligned}\hat{W} &= \arg\max_W \tilde{p}(W|\mathbf{X}) \\ &= \arg\max_W \tilde{p}(\mathbf{X}|W) \cdot P_{\Gamma_0}(W)\end{aligned} \tag{74}$$

where

$$\tilde{p}(\mathbf{X}|W) = \int p(\mathbf{X}|\Lambda, W)p(\Lambda|\varphi)\, d\Lambda \tag{75}$$

is called the *predictive pdf* (e.g., [3], [57], and [142]) of the observation $\mathbf{X}$ given the word W. The crucial difference between the plug-in and predictive classifiers is that the former acts as if the estimated model parameters were the true ones, whereas the predictive methods average over the uncertainty in parameters. Three key issues thus arise in BPC:

1) the definition of the prior density $p(\Lambda|\varphi)$ for modeling the uncertainty of the HMM parameters;
2) the specification of the hyperparameters, φ;
3) the evaluation of the predictive density.

Readers are referred to [74], [75], [83], and [84] for details on how the above issues are addressed in a series of preliminary studies and how the BPC approach enhances robustness when mismatches exist between training and testing conditions.

C. Related Robust Decision Approaches

If training data can be incorporated into designing decision rules, some new possibility opens. One such example is the *approximate Bayesian* (AB) *decision rule* for speech recognition, which was based on the generalized likelihood ratios computed from the available training and testing data. Such an AB rule operates as follows [121]:

$$\hat{W} = \arg\max_W \frac{\max_\Lambda [p(\mathbf{X}|\Lambda, W) \cdot p(\mathcal{X}|\Lambda, W)]}{\max_\Lambda p(\mathcal{X}|\Lambda, W)]} P_{\Gamma_0}(W). \tag{76}$$

As discussed previously, the minimax classification rule can be viewed as a two-step procedure and implemented in (69). First, each testing utterance is treated as possibly belonging to any word sequence, and a constrained ML estimate of the related HMM parameters is obtained. Then, a plug-in MAP rule is used for speech recognition by using the updated HMM parameters. We can use another estimation technique in the first step and end up with a modified minimax decision rule, e.g.,

$$\hat{W} = \arg\max_W [p(\mathbf{X}|\Lambda_{\mathrm{MAP}}, W) \cdot P_{\Gamma_0}(W)] \tag{77}$$

where Λ_{MAP} is an MAP estimate, $\Lambda_{\mathrm{MAP}} = \arg\max_{\Lambda \in \Omega_\Lambda} p(\mathbf{X}|\Lambda, W)p(\Lambda|\varphi)$. For the convenience of reference, we call this modified minimax decision rule a *Bayesian minimax rule* to emphasize its difference from the original minimax approach in [122]. The readers are referred to [82] for a performance comparison of different implementations of the minimax rule.

We have previously discussed BPC approach as a new decision rule that averages out the sampling error in HMM parameter estimation. A related but simpler approach can also be used for model compensation and adaptation. By assuming the CDHMM and/or transformation parameters to be uncertain, Bayesian predictive densities can be computed for a subset of the parameters. In [149], such an idea is explored in the context of Bayesian speaker adaptation where a Gaussian prior pdf for the mean vector is adopted and the *Bayesian predictive density* of each Gaussian mixture component is calculated to serve as the compensated distribution of that component which is used in the plug-in MAP decision rule in (2). In [83], a similar idea is applied to noisy speech recognition where a uniform prior pdf on a prespecified uncertainty neighborhood for the mean vector is adopted. The *Bayesian predictive compensation* [162] and

Bayesian predictive adaptation [163] are designed to handle a small number of transformation parameters instead of the entire set of CDHMM parameters. Both techniques were found to be robust to speaker and channel distortions when a small size adaptation set was used.

IX. Discussion and Conclusion

We have revisited the classical Bayes' decision theory and discussed how it has been used to design pattern-recognition decision rules such as an automatic speech-recognition algorithm. Due to the lack of a complete knowledge of the joint distribution of patterns and classes in practical pattern-recognition problems, a designer usually assumes a particular form of a parametric distribution and estimates the parameters needed to evaluate the joint distribution from a collection of labeled training data. An adaptive decision rule, such as the plug-in maximum *a posteriori* decision rule, is then adopted to perform the desired pattern-recognition operation. We have explained several key concepts about the optimal decision rule, plug-in decision rule, and robust decision rule. We have shown how these decision rules can be derived under different assumptions and optimality criteria. A clear understanding of these aspects will guide us to appreciate why the current ASR technology is so successful in certain applications, and more important, why it fails in many other situations. Although ASR is chosen as the application discussed in this article, we deliberately make our discussions as general as possible so that most of them can be applied to other pattern-recognition problems employing the same decisiontheoretic formulation.

After a careful review of the theoretic foundations of the modern ASR technology, it is quite clear that in order to design an automatic speech recognizer that works well for different tasks and speakers over unexpected and possibly adverse conditions, all of the three distortion types, namely, the small sample effect, the training model and estimation errors, and the mismatched testing conditions, discussed in Section II-D, need to be appropriately treated to deal with these violations of modeling assumptions. Not all of them has been seriously addressed in the past. In this paper, we have mainly addressed issues related to adaptive modeling of speech and linguistic units. We have also briefly discussed a recent research trend in designing some new robust decision rules. These rules will be especially attractive for the class of robust speech-recognition problem in which:

1) mismatches between training and testing conditions exist;
2) an accurate knowledge of the mismatch mechanism is unknown;
3) the only available information is the test data along with a set of pretrained speech models and the decision parameters.

More fundamental work and research innovations are needed in this area.

Before we close this paper, we want to emphasize again that in order to derive an optimal decision rule, it is important to have correct knowledge of three key factors, namely, the observation space Ω_x, the loss function $\ell(W,\, d(\mathbf{X}))$, and the joint pdf $p(W,\, \mathbf{X})$. Based on the previous discussion, it is quite clear to us now that the performance of the currently popular ASR systems, which adopt a plug-in MAP decision rule with ML/MAP-estimated densities, will depend on the following conditions:

1) whether the assumed parametric models are accurate and flexible enough to appropriately model the highly complex and variable speech signals or the extracted feature vectors;
2) whether the training data set is sufficient and representative enough to guarantee good parameter estimation and generalizability;
3) whether the assumed models and the related parameter estimation methods are computationally efficient and robust enough to take care of the possible distortions between models and training samples;
4) whether the distortions between the trained models and the actual testing data are small enough to avoid the breakdown of the whole approach.

We can always try to improve the ASR performance by:

1) finding invariant or robust speech features (i.e., a better Ω_x);
2) developing better modeling and learning techniques [i.e., a better $p(W,\, \mathbf{X})$];
3) applying adaptation techniques [i.e., a better $p(W,\, \mathbf{X})$];
4) using robust decision strategies (i.e., try to make the best decision based on all of the available information).

Among many research issues, we want to emphasize the importance of the following issues:

1) how to collect/find useful real speech data;
2) how to efficiently and intelligently use these training data to discover useful knowledge sources;
3) how to use the above derived knowledge sources in designing a robust ASR system;
4) how to incorporate confidence measures into recognized words and phrases to improve *intelligence* of speech-recognition systems.

Technical advances are needed in discovering new structures in signal, feature, and model representations and their interactions with the speaker and speaking environment in which the speech signal is generated. This will allow us to incorporate more knowledge sources in the source-channel models shown in Fig. 2 to more faithfully reflect the actual channel information illustrated in Fig. 1. Our discussion in this paper about model estimation, adaptation, compensation, and normalization coupled with these new advances will guide us in designing high-performance and robust decision rules in the future. It is our hope that the above in-depth discussions may inspire further innovations that will lead to better solutions for automatic speech recognition and many other pattern-recognition problems.

Acknowledgment

The authors gratefully acknowledge the contributions of many of their past and present collaborators, including C.

Chan, C. Chesta, J.-T. Chien, W. Chou, J.-L. Gauvain, H. Jiang, B.-H. Juang, S. Katagiri, C.-H. Lin, B. Ma, T. Matsuoka, N. Merhav, T. A. Myrvoll, L. R. Rabiner, M. Rahim, A. Sankar, K. Shinoda, O. Siohan, and A. C. Surendran. Their insight and hard work have made the topics of decision rule design and decision parameter adaptation two of the most fruitful areas in the field of automatic speech recognition in recent years. The materials presented in this paper are largely extracted from some of their joint publications or discussions with the authors.

The authors also thank O. Siohan for sharing with them the initial versions of the roadmap in Fig. 3 and the illustration of direct and indirect HMM adaptation in Fig. 4. Both figures were later updated by the authors to reflect recent advances of parameter adaptation and joint estimation of transformation and HMM parameters for speech recognition. The authors also owe their appreciation to two anonymous reviewers, whose comments helped the presentation of this paper.

REFERENCES

[1] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*. Norwell, MA: Kluwer, 1993.

[2] S. M. Ahadi and P. C. Woodland, "Combined Bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 11, pp. 187–206, 1997.

[3] J. Aitchison and I. R. Dunsmore, *Statistical Prediction Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1975.

[4] S. Amari, "A theory of adaptive pattern classifiers," *IEEE Trans. Electron. Comput.*, vol. EC-16, no. 3, pp. 299–307, 1967.

[5] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker adaptive training," in *Proc. ICSLP-96*, Philadelphia, PA, 1996, pp. 1137–1140.

[6] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304–1312, 1974.

[7] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, no. 2, pp. 179–190, 1983.

[8] 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. ICASSP-86*, Tokyo, Japan, 1986, pp. 49–52.

[9] ——, "Tree-based language model for natural language speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1001–1008, July 1989.

[10] L. R. Bahl, P. V. De Souza, P. S. Gopalakrishnan, and M. A. Picheny, "Context dependent vector quantization for continuous speech recognition," in *Proc. ICASSP-93*, Minneapolis, MN, 1993, pp. II-632–II-635.

[11] L. R. Bahl, P. F. Brown, P. V. De Souza, and R. L. Mercer, "Estimating hidden Markov model parameters so as to maximize speech recognition accuracy," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 1, pp. 77–83, 1993.

[12] J. K. Baker, "Stochastic modeling for automatic speech understanding," in *Speech Recognition*, D. R. Reddy, Ed. New York: Academic, 1975, pp. 521–542.

[13] ——, "The Dragon system—An overview," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 24–29, Jan. 1975.

[14] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, vol. 41, pp. 164–171, 1970.

[15] L. E. Baum, "An inequality and associated maximization techniques in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1–8, 1972.

[16] J. R. Bellegarda and D. Nahamoo, "Tied mixture continuous parameter modeling for speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 2033–2045, Dec. 1990.

[17] J. R. Bellegarda, P. V. De Souza, A. J. Nadas, D. Nahamoo, M. A. Picheny, and L. R. Bahl, "The metamorphic algorithm: A speaker mapping approach to data augmentation," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 413–420, July 1994.

[18] J. R. Bellegarda, "A multispan statistical language modeling for large vocabulary speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 456–467, Sept. 1998.

[19] ——, "Exploiting latent semantic information in statistical language modeling," *Proc. IEEE*, vol. 88, pp. 1279–1296, Aug. 2000.

[20] J. M. Bernardo and F. J. Giron, "A Bayesian analysis of simple mixture problems," in *Bayesian Statistics 3*, J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Eds. Oxford, U.K.: Oxford Univ. Press, 1988, pp. 67–78.

[21] B. P. Carlin and T. A. Louis, *Bayes and Empirical Bayes Methods for Data Analysis*. London, U.K.: Chapman & Hall, 1996.

[22] S. Chen and P. V. De Souza, "Speaker adaptation by correlation (ABC)," in *Proc. DARPA SLT Workshop*, 1997.

[23] J.-T. Chien and H.-C. Wang, "Telephone speech recognition based on Bayesian adaptation of hidden Markov models," *Speech Commun.*, vol. 22, pp. 369–384, 1997.

[24] J.-T. Chien, C.-H. Lee, and H.-C. Wang, "A hybrid algorithm for speaker adaptation using MAP transformation and adaptation," *IEEE Signal Processing Lett.*, vol. 4, pp. 167–168, June 1997.

[25] ——, "Improved Bayesian learning of hidden Markov models for speaker adaptation," in *Proc. ICASSP-97*, Munich, Germany, 1997, pp. 1027–1030.

[26] J.-T. Chien, "On-line hierarchical transformation of hidden Markov models for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 656–667, Nov. 1999.

[27] W. Chou, B.-H. Juang, and C.-H. Lee, "Segmental GPD training of HMM based speech recognizer," in *Proc. ICASSP-92*, San Francisco, CA, 1992, pp. 473–476.

[28] W. Chou, C.-H. Lee, and B.-H. Juang, "Minimum error rate training based on the N-best string models," in *Proc. ICASSP-93*, Minneapolis, MN, 1993, pp. II-652–II-655.

[29] ——, "Minimum error rate training of inter-word context dependent acoustic model units in speech recognition," in *Proc. ICSLP-94*, Yokohama, Japan, 1994, pp. 439–442.

[30] W. Chou, "Maximum posterior linear regression with elliptically symmetric matrix variate priors," in *Proc. EuroSpeech-99*, Budapest, Hungary, 1999, pp. 1–4.

[31] S. J. Cox and J. S. Bridle, "Unsupervised speaker adaptation by probabilistic fitting," in *Proc. ICASSP-89*, Glasgow, U.K., 1989, pp. 294–297.

[32] S. J. Cox, "Predictive speaker adaptation in speech recognition," *Comput. Speech Lang.*, vol. 9, pp. 1–17, 1995.

[33] M. DeGroot, *Optimal Statistical Decisions*. New York: McGraw-Hill, 1970.

[34] S. A. Della Pietra, V. J. Della Pietra, R. L. Mercer, and S. Roukos, "Adaptive language modeling using minimum discriminant estimation," in *Proc. ICASSP-92*, San Francisco, CA, 1992, pp. 633–636.

[35] L. Delphin-Poulat, C. Mokbel, and J. Idier, "Frame-synchronous stochastic matching based on the Kullback–Leibler information," in *Proc. ICASSP-98*, Seattle, WA, 1998, pp. 89–92.

[36] R. De Mori, Ed., *Spoken Dialogues with Computers*. New York: Academic, 1998.

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

[38] L. Deng, "A dynamic feature based approach to the interface between phonology and phonetics for speech modeling and recognition," *Speech Commun.*, vol. 24, no. 4, pp. 299–323, 1998.

[39] V. V. Digalakis, D. Ritchev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 357–366, Sept. 1995.

[40] V. V. Digalakis and L. G. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 294–300, July 1996.

[41] V. V. Digalakis, "Online adaptation of Hidden Markov models using incremental estimation algorithms," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 253–261, May 1999.

[42] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[43] Y. Ephraim, A. Dembo, and L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1001–1013, Sept. 1989.

[44] Y. Ephraim and L. R. Rabiner, "On the relations between modeling approaches for speech recognition," *IEEE Trans. Inform. Theory*, vol. 36, pp. 372–380, Mar. 1990.

[45] Y. Ephraim, "Statistical model based speech enhancement systems," *Proc. IEEE*, vol. 80, pp. 1526–1555, Oct. 1992.

[46] M. Federico, "Bayesian estimation method of N-gram language model adaptation," in *Proc. ICSLP-96*, Philadelphia, PA, 1996, pp. 240–243.

[47] J. Ferguson, Ed., *Hidden Markov Models for Speech*. Princeton, NJ: IDA, 1980.

[48] T. S. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic, 1967.

[49] S. Furui, "A training procedure for isolated word recognition systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 2, pp. 129–136, 1980.

[50] ——, "Unsupervised speaker adaptation method based on hierarchical spectral clustering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1923–1930, Dec. 1989.

[51] M. J. F. Gales and S. J. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Comput. Speech Lang.*, vol. 9, pp. 289–307, 1995.

[52] ——, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 352–359, Sept. 1996.

[53] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR Framework," *Comput. Speech Lang.*, vol. 10, pp. 249–264, 1996.

[54] Y.-Q. Gao, M. Padmanabhan, and M. A. Picheny, "Speaker adaptation based on pre-clustering training speakers," in *Proc. EuroSpeech-97*, Rhodes, Greece, 1997, pp. 2091–2094.

[55] J.-L. Gauvain and C.-H. Lee, "Bayesian learning for hidden Markov models with Gaussian mixture state observation densities," *Speech Commun.*, vol. 11, no. 2–3, pp. 205–214, 1992.

[56] ——, "Maximum A posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291–298, Apr. 1994.

[57] S. Geisser, *Predictive Inference: An Introduction*. New York: Chapman & Hall, 1993.

[58] N. Glick, "Sample-based classification procedures derived from density estimators," *J. Amer: Statist. Assoc.*, vol. 67, pp. 116–122, 1972.

[59] ——, "Sample-based classification procedures related to empiric distributions," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 454–461, 1976.

[60] I. J. Good, "The population frequencies species and the estimation of population parameters," *Biometrika*, vol. 40, pp. 237–264, 1953.

[61] Y. Gong, "Stochastic trajectory modeling and sentence searching for continuous speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 33–44, Jan. 1997.

[62] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas, D. Nahamoo, and M. A. Picheny, "Decoder selection based on cross-entropies," in *Proc. ICASSP-88*, New York, 1988, pp. 20–23.

[63] A. K. Gupta and T. Varga, *Elliptically Contoured Models in Statistics*. Norwell, MA: Kluwer, 1993.

[64] J. D. Hamilton, "A quasi-Bayesian approach to estimating parameters for mixtures of normal distributions," *J. Bus. Econ. Statist.*, vol. 9, no. 1, pp. 27–39, 1991.

[65] H. Hattori and S. Sagayama, "Vector field smoothing principle for speaker adaptation," in *Proc. ICSLP-92*, Banff, Alberta, Canada, 1992, pp. 381–384.

[66] H.-W. Hon, "Vocabulary-independent speech recognition: The VOCIND System," Ph.D. dissertation, School of Comput. Sci., Carnegie-Mellon Univ., Pittsburgh, PA, 1992.

[67] X. Huang and M. A. Jack, "Semi-continuous hidden Markov models for speech signal," *Comput. Speech Lang.*, vol. 3, no. 3, pp. 239–251, 1989.

[68] X. Huang and K.-F. Lee, "On speaker-independent speaker dependent and speaker-adaptive speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 150–157, Apr. 1993.

[69] Q. Huo, C. Chan, and C.-H. Lee, "Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 334–345, Sept. 1995.

[70] ——, "On-line adaptation of the SCHMM parameters based on the segmental quasi-Bayes learning for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 141–144, Mar. 1996.

[71] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 161–172, Mar. 1997.

[72] ——, "On-line adaptive learning of the correlated continuous density hidden Markov models for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 386–397, July 1998.

[73] Q. Huo and B. Ma, "On-line adaptive learning of CDHMM parameters based on multiple-stream prior evolution and posterior pooling," in *Proc. EuroSpeech-99*, Budapest, Hungary, 1999, pp. 2721–2724.

[74] Q. Huo and C.-H. Lee, "A Bayesian predictive classification approach to robust speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 200–204, Mar. 2000.

[75] ——, "Robust speech recognition based on adaptive classification and decision strategies," Speech Commun., to be published.

[76] Q. Huo, N. Smith, and B. Ma, "Efficient ML training of CDHMM parameters based on prior evolution posterior intervention and feedback," in *Proc. ICASSP-2000*, Turkey, to be published.

[77] M.-Y. Hwang, "Subphonetic acoustic modeling for speaker-independent continuous speech recognition," Ph.D. dissertation, School of Comput. Sci., Carnegie-Mellon Univ., Pittsburgh, PA, 1993.

[78] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, pp. 532–556, Apr. 1976.

[79] F. Jelinek, R. L. Mercer, and S. Roukos, "Principles of lexical language modeling for speech recognition," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1991, pp. 651–699.

[80] F. Jelinek, B. Merialdo, S. Roukos, and M. Strauss, "A dynamic language model for speech recognition," in *Proc. DARPA Speech and Natural Language Workshop*, Pacific Grove, CA, 1991, pp. 293–295.

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

[82] H. Jiang, K. Hirose, and Q. Huo, "A minimax search algorithm for CDHMM based robust continuous speech recognition," in *Proc. ICSLP-98*, Sydney, Australia, 1998, pp. II-389–II-392.

[83] ——, "Robust speech recognition based on a Bayesian prediction approach," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 426–440, July 1999.

[84] ——, "Improving Viterbi Bayesian predictive classification via sequential Bayesian learning in robust speech recognition," *Speech Commun.*, vol. 28, no. 4, pp. 313–326, 1999.

[85] B.-H. Juang, S. E. Levinson, and M. M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Trans. Inform. Theory*, vol. IT-32, no. 2, pp. 307–309, 1986.

[86] B.-H. Juang and L. R. Rabiner, "The segmental K-means algorithm for estimating parameters of hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1639–1641, Sept. 1990.

[87] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing*, vol. 40, pp. 3043–3054, Dec. 1992.

[88] B.-H. Juang, "Automatic speech recognition: problems progress & prospects," presented at the *1996 IEEE Workshop on Neural Networks For Signal Processing*, Kyoto, Japan, 1996.

[89] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum Classification Error Rate Methods for Speech Recognition," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 257–265, May 1997.

[90] J.-C. Junqua and J.-P. Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. Norwell, MA: Kluwer, 1996.

[91] S. Katagiri, C.-H. Lee, B.-H. Juang, and T. Komori, "New discriminative training algorithms based on a generalized probabilistic descent method," presented at the *Proc. IEEE-SP Workshop Neural Networks for Signal Processing*, Princeton, NJ, 1991.

[92] S. Katagiri, B.-H. Juang, and C.-H. Lee, "Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method," *Proc. IEEE*, vol. 86, pp. 2345–2373, Nov. 1998.

[93] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 400–401, Mar. 1987.

[94] R. Kneser, J. Peters, and D. Klakow, "Language model adaptation using dynamic marginals," in *Proc. EuroSpeech-97*, Rhodes, Greece, 1997, pp. 1971–1974.

[95] Y. Kharin, *Robustness in Statistical Pattern Recognition*. Norwell, MA: Kluwer, 1996.

[96] V. Krishnamurthy and J. B. Moore, "On-line estimation of Hidden Markov model parameters based on the Kullback–Leibler Information measure," *IEEE Trans. Signal Processing*, vol. 41, pp. 2557–2573, Aug. 1993.
[97] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.
[98] R. Kuhn and R. De Mori, "A cache-based natural language model for speech recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 570–583, June 1990.
[99] M. J. Lasry and R. M. Stern, "*A posteriori* estimation of correlated jointly Gaussian mean vectors," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 530–535, Apr. 1984.
[100] R. Lau, R. Rosenfield, and S. Roukos, "Trigger-based language models: A maximum entropy approach," in *Proc. ICASSP-93*, Minneapolis, MN, 1993, pp. II-45–II-48.
[101] C.-H. Lee, L. R. Rabiner, R. Pieraccini, and J. G. Wilpon, "Acoustic modeling for large vocabulary speech recognition," *Comput. Speech Lang.*, vol. 4, pp. 127–165, 1990.
[102] C.-H. Lee, C.-H. Lin, and B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. Signal Processing*, vol. 39, pp. 806–814, Apr. 1991.
[103] C.-H. Lee and J.-L. Gauvain, "Speaker adaptation based on MAP estimation of HMM parameters," in *Proc. ICASSP-93*, Minneapolis, MN, 1993, pp. II-652–II-655.
[104] C.-H. Lee, F.-K. Soong, and K.-K. Paliwal, Eds., *Automatic Speech and Speaker Recognition: Advanced Topics*. Norwell, MA: Kluwer, 1996.
[105] C.-H. Lee, B.-H. Juang, W. Chou, and J. J. Molina-Perez, "A study on task-independent subword selection and modeling for speech recognition," in *Proc. ICSLP-96*, Philadelphia, PA, 1996, pp. 1816–1819.
[106] C.-H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition," *Speech Commun.*, vol. 25, pp. 29–47, 1998.
[107] K.-F. Lee, *Automatic Speech Recognition—The Development of the SPHINX-System*. Norwell, MA: Kluwer, 1989.
[108] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP-96*, Atlanta, GA, 1996, pp. 353–356.
[109] 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, pp. 171–185, 1995.
[110] ——, "Flexible speaker adaptation using maximum likelihood linear regression," in *Proc. ARPA SLS Technology Workshop*, 1995, pp. 110–115.
[111] S. E. Levinson, "Structural methods in automatic speech recognition," *Proc. IEEE*, vol. 73, pp. 1625–1650, 1985.
[112] L. R. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-28, no. 5, pp. 729–734, 1982.
[113] C.-S. Liu, C.-H. Lee, W. Chou, A. E. Rosenberg, and B.-H. Juang, "A study on minimum error discriminative training for speaker recognition," *J. Acoust. Soc. Amer.*, vol. 97, no. 1, pp. 637–648, 1995.
[114] F.-H. Liu, "Environment adaptation for robust speech recognition," Ph.D. dissertation, School of Comput. Sci. Carnegie-Mellon Univ., Pittsburgh, PA, 1994.
[115] A. Ljolje, Y. Ephraim, and L. R. Rabiner, "Estimation of hidden Markov model parameters by minimizing empirical error rate," in *Proc. ICASSP-90*, 1990, pp. 709–712.
[116] U. E. Makov and A. F. M. Smith, "A quasi-Bayes unsupervised learning procedure for priors," *IEEE Trans. Inform. Theory*, vol. IT-23, no. 6, pp. 761–764, 1977.
[117] J. S. Maritz and T. Lwin, *Empirical Bayes Methods*, 2nd ed. New York: Chapman & Hall, 1989.
[118] F. Martin, K. Shikano, and Y. Minami, "Recognition of noisy speech by composition of hidden Markov models," in *Proc. EuroSpeech-93*, Berlin, Germany, 1993, pp. 1031–1034.
[119] J. J. Martin, *Bayesian Decision Problems and Markov Chains*. New York: Wiley, 1967.
[120] T. Matsuoka and C.-H. Lee, "A study of on-line Bayesian adaptation for HMM-based speech recognition," in *Proc. EuroSpeech-93*, Berlin, Germany, 1993, pp. 815–818.
[121] N. Merhav and Y. Ephraim, "A Bayesian classification approach with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 2157–2166, Oct. 1991.
[122] N. Merhav and C.-H. Lee, "A minimax classification approach with application to robust speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 90–100, Jan. 1993.
[123] C. Mokbel and L. Delphin-Poulat, "A Unified framework for auto-adaptive speech recognition," in *Proc. Workshop Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, 1999, pp. 227–230.
[124] S. Moon and J.-N. Hwang, "Robust speech recognition based on joint model and feature space optimization of hidden Markov models," *IEEE Trans. Neural Networks*, vol. 8, pp. 194–204, Mar. 1997.
[125] A. Nadas, "A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 814–817, Apr. 1983.
[126] ——, "Optimal solution of a training problem in speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 326–329, Jan. 1985.
[127] A. Nadas, D. Nahamoo, and M. A. Picheny, "On a Model-robust training method for speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1432–1436, Sept. 1988.
[128] K. Ohkura, M. Sugiyama, and S. Sagayama, "Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMMs," in *Proc. ICSLP-92*, Banff, Alberta, Canada, 1992, pp. 369–372.
[129] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, "From HMMs to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 360–378, Sept. 1996.
[130] M. Padmanabhan, L. R. Bahl, D. Nahamoo, and M. A. Picheny, "Speaker Clustering and transformation for speaker adaptation in speech recognition systems," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 71–77, Jan. 1998.
[131] P. Placeway, R. Schwartz, P. Fung, and L. Nguyen, "The estimation of powerful language models from small and large corpora," in *Proc. ICASSP-93*, vol. 2, Minneapolis, MN, 1993, pp. 33–36.
[132] L. R. Rabiner, J. G. Wilpon, and B.-H. Juang, "A segmental K-means training procedure for connected word recognition," *AT&T Tech. J.*, vol. 65, pp. 21–31, 1986.
[133] ——, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
[134] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[135] M. Rahim and B.-H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 19–30, Jan. 1996.
[136] M. Rahim and C.-H. Lee, "Simultaneous feature and HMM design using string-based minimum classification error training criterion," in *Proc. ICSLP-96*, Philadelphia, PA, 1996, pp. 1820–1823.
[137] M. Rahim, C.-H. Lee, and B.-H. Juang, "Discriminative utterance verification for connected digit recognition," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 266–277, May 1997.
[138] C. Rathinavelu and L. Deng, "Use of generalized dynamic feature parameters for speech recognition: Maximum likelihood and minimum classification error approaches," in *Proc. ICASSP-95*, Detroit, MI, 1995, pp. 373–376.
[139] ——, "Speaker adaptation experiments using nonstationary-state hidden Markov models: A MAP approach," in *Proc. ICASSP-97*, Munich, Germany, 1997, pp. 1415–1418.
[140] W. Reichl, "Language model adaptation using minimum discrimination information," in *Proc. EuroSpeech-99*, Budapest, Hungary, 1999, pp. 1791–1794.
[141] R. A. Redner and H. F. Walker, "Mixture densities maximum likelihood and the EM algorithm," *SIAM Rev.*, vol. 26, no. 2, pp. 195–239, 1984.
[142] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge, U.K.: Cambridge Univ. Press, 1996.
[143] H. Robbins, "The empirical Bayes approach to statistical decision problems," *Ann. Math. Statist.*, vol. 35, pp. 1–20, 1964.
[144] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of speech and background with application to speaker identification in noise," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 245–257, Apr. 1994.
[145] R. Rosenfeld, "Adaptive statistical language modeling: A maximum entropy approach," Ph.D. dissertation, School of Comput. Sci., Carnegie-Mellon Univ., Pittsburgh, PA, 1994.

[146] ——, "Two decades of statistical language modeling: Where do we go from here?," *Proc. IEEE*, vol. 88, pp. 1270–1278, Aug. 2000.
[147] A. Sankar and C.-H. Lee, "A maximum likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 190–202, May 1996.
[148] R. Schwartz and F. Kubala, "Hidden Markov models and speaker adaptation," in *Speech Recognition and Understanding—Recent Advances Trends and Applications*, ser. NATO ASI F75, P. Laface and R. De Mori, Eds., 1991, pp. 31–57.
[149] B. M. Shahshahani, "A Markov random field approach to Bayesian speaker adaptation," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 2, pp. 183–191, 1997.
[150] K. Shinoda and T. Watanabe, "Speaker adaptation with autonomous model complexity control by MDI principle," in *Proc. ICASSP-96*, Atlanta, GA, 1996, pp. 717–720.
[151] K. Shinoda and C.-H. Lee, "Structural MAP speaker adaptation using hierarchical priors," in *Proc. 1997 IEEE Workshop Automatic Speech Recognition and Understanding*, Santa Barbara, CA, 1997, pp. 381–388.
[152] ——, "Unsupervised adaptation using structural Bayes approach," in *Proc. ICASSP-98*, Seattle, WA, 1998, pp. 793–796.
[153] O. Siohan and C.-H. Lee, "Iterative Noise and channel estimation under the stochastic matching algorithm framework," *IEEE Signal Processing Lett.*, vol. 4, pp. 304–306, Nov. 1997.
[154] O. Siohan, C. Chesta, and C.-H. Lee, "Hidden Markov model adaptation using maximum *a posteriori* linear regression," in *Proc. Workshop Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, 1999, pp. 147–150.
[155] ——, "Joint maximum *a posteriori* adaptation of transformation and HMM parameters," in Proc. ICASSP-2000, Turkey, pp. 965–968, to be published.
[156] O. Siohan, T. A. Myrvoll, and C.-H. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation,", submitted for publication.
[157] A. F. M. Smith and U. E. Makov, "A quasi-Bayes sequential procedure for mixtures," *Roy. Statist. Soc. J. B*, vol. 40, no. 1, pp. 106–112, 1978.
[158] "Special issue on Robust speech recognition," *Speech Communication*, vol. 25, no. 1–3, 1998.
[159] J. Spragins, "A note on the iterative application of Bayes' rule," *IEEE Trans. Inform. Theory*, vol. IT-11, no. 4, pp. 544–549, 1965.
[160] R. M. Stern and M. J. Lasry, "Dynamic speaker adaptation for feature-based isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 751–763, June 1987.
[161] R. A. Sukkar and C.-H. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 420–429, Nov. 1996.
[162] A. C. Surendran and C.-H. Lee, "Predictive adaptation and compensation for robust speech recognition," in *Proc. ICSLP-98*, Sydney, Australia, 1998.
[163] ——, "Bayesian predictive approach to adaptation of HMMs," in *Proc. Workshop Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, 1999, pp. 155–158.
[164] A. C. Surendran, C.-H. Lee, and M. Rahim, "Non-linear compensation for stochastic matching," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 643–655, Nov. 1999.
[165] K. Takagi, H. Hattori, and T. Watanabe, "Speech recognition with rapid environment adaptation by spectrum equalization," in *Proc. ICSLP-94*, Yokohama, Japan, 1994, pp. 1023–1026.
[166] J.-I. Takahashi and S. Sagayama, "Vector-field-smoothed Bayesian learning for incremental speaker adaptation," in *Proc. ICASSP-95*, Detroit, MI, 1995, pp. 696–699.
[167] ——, "Tied-Structure HMM based on parameter correlation for efficient model training," in *Proc. ICASSP-96*, Atlanta, GA, 1996, pp. 467–670.
[168] M. Tonomura, T. Kosaka, and S. Matsunaga, "Speaker adaptation based on transfer vector field smoothing using maximum *a posteriori* probability estimation," in *Proc. ICASSP-95*, Detroit, MI, 1995, pp. I-688–I-691.
[169] S. Vaseghi and B. Milner, "A comparative analysis of channel-robust features and channel equalization methods for speech recognition," in *Proc. ICSLP-96*, Philadelphia, PA, 1996, pp. 877–890.
[170] A. Varga and R. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. ICASSP-90*, Albuquerque, NM, 1990, pp. 845–848.
[171] A. Wald, *Statistical Decision Functions*. New York: Wiley, 1950.
[172] S. Wang and Y. Zhao, "On-line tree-structured transformation of hidden Markov models for speaker adaptation," in *Proc. 1999 IEEE Workshop Automatic Speech Recognition and Understanding*, Keystone, 1999.
[173] E. Weinstein, M. Feder, and A. V. Oppenheim, "Sequential algorithms for parameter estimation based on the Kullback–Leibler information measure," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1652–1654, Sept. 1990.
[174] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, "Large vocabulary continuous speech recognition using HTK," in *Proc. ICASSP-94*, Adelaide, 1994, pp. II-125–II-128.
[175] P. C. Woodland, D. Pye, and M. J. F. Gales, "Iterative unsupervised adaptation using maximum likelihood linear regression," in Proc. ICSLP-96, Philadelphia, PA, 1996, pp. 1133–1136, submitted for publication.
[176] ——, "Speaker adaptation: Techniques and challenges," in *Proc. 1999 IEEE Workshop Automatic Speech Recognition and Understanding*, Keystone, 1999.
[177] C. Yen, S.-S. Kuo, and C.-H. Lee, "Minimum error rate training for PHMM-based text recognition," *IEEE Trans. Image Processing*, vol. 8, pp. 1120–1124, Aug. 1999.
[178] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modeling," in *Proc. ARPA Human Language Technology Workshop*, 1994, pp. 307–312.
[179] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 2.1)*. Cambridge, U.K.: Cambridge Univ. Press, 1997.
[180] G. Zavaliagkos, R. Schwartz, and J. Makhoul, "Batch incremental and instantaneous adaptation techniques for speech recognition," in *Proc. ICASSP-95*, Detroit, MI, 1995, pp. I-676–I-679.
[181] Y. Zhao, "An acoustic-phonetic-based speaker adaptation technique for improving speaker-independent continuous speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 380–394, July 1994.
[182] ——, "Self-learning speaker and channel adaptation based on spectral variation source decomposition," *Speech Commun.*, vol. 18, pp. 65–77, 1996.

Chin-Hui Lee (Fellow, IEEE) received the B.S. degree from National Taiwan University, Taipei, in 1973, the M.S. degree from Yale University, New Haven, CT, in 1977, and the Ph.D. degree from the University of Washington, Seattle, in 1981, all in electrical engineering.

In 1981, he joined Verbex Corporation, Bedford, MA, and was involved in research work on connected word recognition. In 1984, he became affiliated with Digital Sound Corporation, Santa Barbara, CA, where he engaged in research in speech coding, speech recognition, and signal processing for the development of the DSC-2000 Voice Server. Since 1986, he has been with Bell Laboratories, Murray Hill, NJ, where he is currently a Distinguished Member of Technical Staff and Head of Dialogue Systems Research Department. He has published more than 200 papers in journals and international conferences and workshops on the topics in automatic speech and speaker recognition. His research scope is reflected in an edited book, *Automatic Speech and Speaker Recognition: Advanced Topics* (Norwell, MA: Kluwer, 1996). His current research interests include multimedia signal processing, speech and language modeling, adaptive and discriminative modeling, speech recognition, speaker recognition, spoken dialogue processing, biometrics, and human–machine interface.

Dr. Lee is a recipient of the 1994 SPS Senior Award and the 1997 and 1999 SPS Best Paper Award in Speech Processing. He was a winner of the prestigious Bell Laboratories President Gold Award in 1997 for his contributions to the Bell Labs Automatic Speech Recognition algorithms and products. Recently, he was named as one of the six distinguished lecturers of SPS for the year 2000. From 1991 to 1995, he was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He was a member of the ARPA Spoken Language Coordination Committee during the same period. Since 1995, he has been a member of the Speech Processing Technical Committee of the IEEE Signal Processing Society (SPS), in which he served as Chairman from 1996 to 1998. In 1996, he helped promote the newly formed SPS Multimedia Signal Processing (MMSP) Technical Committee and is now a member.

Qiang Huo (Member, IEEE) received the B.Eng. degree from the University of Science and Technology of China (USTC), Hefei, China, in 1987, the M.Eng. degree from Zhejiang University, Hangzhou, China, in 1989, and the Ph.D. degree from the USTC in 1994, all in electrical engineering.

From 1986 to 1990, his research work focused on the hardware design and development for real-time digital signal processing, image processing and computer vision, speech and speaker recognition. From 1991 to 1994, he was with the Department of Computer Science, The University of Hong Kong (HKU), where he worked on speech recognition. From 1995 to 1997, he was with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan, where he engaged in research in speech recognition. He joined the Department of Computer Science and Information Systems, HKU, again in 1998 as an Assistant Professor. His current major research interests include speech and speaker recognition, computational model for spoken dialogue processing, Chinese character recognition, biometric authentication, adaptive signal modeling and processing, and general pattern recognition theory.