Chapter 4
Front-End, Back-End, and Hybrid Techniques for Noise-Robust Speech Recognition

Li Deng

Abstract Noise robustness has long been an active area of research that captures significant interest from speech recognition researchers and developers. In this chapter, with a focus on the problem of uncertainty handling in robust speech recognition, we use the Bayesian framework as a common thread for connecting, analyzing, and categorizing a number of popular approaches to the solutions pursued in the recent past. The topics covered in this chapter include 1) Bayesian decision rules with unreliable features and unreliable model parameters; 2) principled ways of computing feature uncertainty using structured speech distortion models; 3) use of a phase factor in an advanced speech distortion model for feature compensation; 4) a novel perspective on model compensation as a special implementation of the general Bayesian predictive classification rule capitalizing on model parameter uncertainty; 5) taxonomy of noise compensation techniques using two distinct axes, feature vs. model domain and structured vs. unstructured transformation; and 6) noise-adaptive training as a hybrid feature-model compensation framework and its various forms of extension.

4.1 Introduction

Noise-robust speech recognition has been an active area of research for many years, and is still a vigorous research area with many practical applications today, e.g., [1, 7–10, 17, 18, 24, 39, 59, 66, 80, 85]. There are numerous challenges to building a speech recognition system that is robust to environmental noise. Noise is unpredictable, time-varying, and has a variety of properties. Not only is accurate noise estimation itself a difficult problem, but, even given an accurate model of noise, nonlinear interactions between clean speech and noise in generating noise-distorted speech (often parameterized in the log power spectrum or cepstrum) also give rise to high complexity in decoding speech, with a high degree of imperfection.

Microsoft Research, One Microsoft Way, Redmond, WA 98052
deng@microsoft.com

Standard noise robustness methods can be divided into the broad categories of feature compensation and model compensation. Feature compensation is also called feature enhancement or front-end denoising, where the effect of noise is removed from the observed noisy speech features without using speech decoding results and without changing the parameters of the acoustic model (e.g., HMM). On the other hand, in model compensation the acoustic model parameters can be modified to incorporate the effects of noise, where each component of the model can be adapted to account for how the noise affects its mean and variance. While typically achieving higher performance than feature compensation, model compensation often incurs significantly greater computational cost with straightforward implementation (unless drastic approximations are made, such as the use of sub-space techniques, e.g., [64]).

In the past decade, with the earliest work published in the same year [5,19,32,57], a new approach to robust ASR has emerged that is aimed at propagating the uncertainty in the acoustic features due to either the noise effect itself or the residual noise after feature compensation into the decoding process of speech recognition. These techniques provide, at the frame level, dynamic compensation of HMM variances as well as its means, based on the estimate of uncertainty caused by imperfect feature enhancement, and incorporate the compensated parameters into the decoding process. The goal is to achieve recognition performance that is comparable to model compensation techniques, with computational cost similar to feature compensation. This way of handling uncertainty in speech feature data, the theme of this book, appears to strike a balance between computational cost and noise-robust speech recognition accuracy. Several key issues related to the scheme of the above uncertainty decoding have been addressed by a number of researchers, with various new approaches proposed and developed in [28,43,50,51,65–67,82].

Given the large number of noise-robust speech recognition techniques developed over the past two decades, this chapter provides a selective overview of them and focuses on the topics that have particular relevance for the future development of noise-robust speech recognition technology. Section 2 starts by introducing the Bayesian perspective as a common thread that connects the remaining topics presented in this chapter, and provides a theoretical background for treating uncertain data. A concrete example is given in Section 3 on how data uncertainty produced in feature compensation can be computed in a principled way. The Algonquin model of speech distortion is used in the example. A more detailed model of speech distortion and mismatch than Algonquin, which makes use of the non-uniform distribution of phase factors between clean speech and mixing noise (due to many data points in filter banks while computing cepstral features), is presented and discussed in Section 4. Feature compensation experiments using this phase-sensitive model and insights gained from these experiments are also provided. In Section 5, we shift the discussion from feature-domain compensation and the associated uncertainty to their model-domain counterparts. Uncertainty in model parameters is studied within the Bayesian framework, where both feature and model uncertainties are integrated into a most general form of Bayesian predictive classification rule. Importantly, in this general framework, model compensation can be viewed as a special realiza-

tion of the Bayesian predictive classification rule. Then, in Section 6, a taxonomy of a multitude of model compensation techniques is provided based on the structured vs. unstructured transformation of the model parameters. This same axis of structured vs. unstructured transformation in the feature domain is also used to categorize a multitude of feature compensation techniques. Finally, in Section 7, we discuss hybrid feature-model compensation techniques and, in particular, its important member, noise-adaptive training.

4.2 Bayesian Decision Rule with Unreliable Features

In this section, we start from the Bayesian framework to account for uncertainty of the features that form the input into speech recognition systems. This framework provides the theoretical basis for "uncertainty decoding", which manifests itself in various forms in the literature. In Section 5 of this chapter, we apply the same framework to account for uncertainty of model parameters, providing the basis for model compensation.

The standard Bayesian decision rule for speech recognition, with fixed model parameter set Λ, gives

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{W},\Lambda)P(\mathbf{W}), \tag{4.1}$$

where $P(\mathbf{W})$ is the prior probability that the speaker utters a word sequence $\mathbf{W}$, and $p(\mathbf{y}|\Lambda,\mathbf{W})$ is the probability that the speaker produces the acoustic feature sequence $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_t, ...\mathbf{y}_T]$ when $\mathbf{W}$ is the intended word sequence. The computation of probability $P(\mathbf{y}|\Lambda,\mathbf{W})$ uses deterministic parameters, denoted by Λ, in the speech model.

Using the rule of total probability and conditional independence, we have

$$p(\mathbf{W},\mathbf{y}) = \int p(\mathbf{W},\mathbf{y},\mathbf{x})d\mathbf{x} = \int p(\mathbf{W}|\mathbf{x})p(\mathbf{x}|\mathbf{y})d\mathbf{x}. \tag{4.2}$$

Thus Eq. (4.1) becomes

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}} \int p(\mathbf{W}|\mathbf{x},\Lambda)p(\mathbf{x}|\mathbf{y})d\mathbf{x}, \tag{4.3}$$

where the conditional $p(\mathbf{x}|\mathbf{y})$ represents the effect of feature compensation from noisy speech $\mathbf{y}$ to enhanced speech $\mathbf{x}$, and $p(\mathbf{W}|\mathbf{x},\Lambda)$ is the objective function for the decoding problem of speech recognition with input feature $\mathbf{x}$.

Applying Bayes' rule on $p(\mathbf{W}|\mathbf{x})$, we obtain from Eq. (4.3)

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}} \int \frac{p(\mathbf{x}|\mathbf{W},\Lambda)}{p(\mathbf{x})} p(\mathbf{x}|\mathbf{y},\Lambda)d\mathbf{x}P(\mathbf{W}). \tag{4.4}$$

Since the prior speech distribution of $p(\mathbf{x})$ is sufficiently broad, with its variance being significantly larger than the posterior (an assumption which was also used in [32]), we can simplify the rule by assuming it is approximately constant over the range of $p(\mathbf{x})$ values of interest. Thus, Eq. (4.4) becomes

$$\hat{\mathbf{W}} \approx \underset{\mathbf{W}}{\operatorname{argmax}} \int p(\mathbf{x}|\mathbf{W},\Lambda)p(\mathbf{x}|\mathbf{y},\Lambda)d\mathbf{x}P(\mathbf{W}), \tag{4.5}$$

which is the uncertainty decoding rule used in [19, 28, 50]. It was pointed out in [66] that under low-SNR conditions the above assumption no longer holds. This observation accounts for poor uncertainty decoding results at low SNR.

One main advantage of the approximate uncertainty decoding rule of Eq. (4.5) is the simplicity in its incorporation into the recognizer's decision rule. This simplicity arises from the fact that a product of two Gaussians remains a Gaussian. One major improvement over the conventional uncertainty decoding rule of Eq. (4.5) is the exploitation of temporal correlations in this rule. The work of [51] explicitly models such correlations in the context of uncertainty decoding. A similar motivation for explicitly exploiting the temporal correlation is presented in [20, 27] in the context of feature compensation, which gives the mean values of the distribution $p(\mathbf{x}|\mathbf{y})$. On the other hand, appending differential parameters to the static ones in the implementation of the uncertainty decoding rule implicitly represents the temporal correlations, as carried out in the work of [28,50]. Another major improvement comes from a series of work reported in [65–67], where the authors developed "joint uncertainty decoding" in which the acoustic space is subdivided into regions and the joint density of clean and noisy speech is estimated using stereo data.

One of most practical issues in uncertainty decoding is the computation of feature uncertainty from $p(\mathbf{x}|\mathbf{y})$, or equivalently $p(\mathbf{y}|\mathbf{x}) = \alpha \frac{p(\mathbf{x}|\mathbf{y})}{p(\mathbf{x})}$, in Eq. (4.4). In [32], the SPLICE model, originally developed in [17, 18] with numerous further improvements [3, 18, 22, 25, 30, 33, 87], was used to compute feature uncertainty based on some approximation of $p(\mathbf{y}|\mathbf{x})$ (although to a lesser degree than the assumption of flat $p(\mathbf{x})$). Feature uncertainty is expressed directly as a function of the SPLICE parameters determined separately using the techniques described in [17, 25]. In [50, 51], feature uncertainty was determined by detailed analysis in an interesting scenario of distributed speech recognition where bit errors or lost speech data packets on IP connections are encountered.

Finally, in [19, 28], feature uncertainty was computed using a parametric statistical model of acoustic distortion, sometimes called the Algonquin model [37, 38], a probabilistic extension of the commonly used deterministic VTS (Vector Taylor Series) model [2, 56, 71]. Because the approach developed in this work has generality, we will review it in some detail in the following section, and point out further potential of this approach.

4.3 Use of Algonquin Model to Compute Feature Uncertainty

In this section, we use the log-spectral features to represent all acoustic data, be it clean speech, noise, noisy speech, or enhanced speech. We provide an iterative solution, where each iteration contains a closed form solution, for the computation of the "uncertainty" expressed as the MMSE estimate and its variance. It is important to point out that when the complex Fourier transform is taken as the feature to represent the acoustic data, a rather different type of analysis and solution can be used [6].

4.3.1 Algonquin Model of Speech Distortion

The Algonquin algorithm [37, 38, 75] was originally developed as a feature compensation technique, and was extended for use in uncertainty decoding in [19, 28]. Underlying the Algonquin algorithm is the Algonquin model which characterizes the relationship among clean speech, corrupting noise, and noisy speech in the log-spectral or the cepstral domain with an SNR-independent modeling-error residual. Without such a modeling residual, the Algonquin model would be the same as the standard VTS model of speech distortion [2, 56, 71]. The residual captures crude properties of the modeling error in deriving the VTS model. The modeling error includes notably the ignorance of phases between clean speech and noise vectors, both in the log domain involving multiple filter bank channels, in producing the noisy speech vector; see the detailed analysis in Section II of [27]. We now provide an overview of the Algonquin model.

Let $\mathbf{y}, \mathbf{x}$, and $\mathbf{n}$ be single-frame vectors of log mel filter energies for noisy speech, clean speech, and additive noise, respectively. These quantities are shown in [27] to satisfy the following relationship when the phase relationship between clean speech and noise is considered:

$$\begin{aligned}\mathbf{y} &= \mathbf{x} + \log\left[(\mathbf{1}+e^{\mathbf{n}-\mathbf{x}})[\mathbf{1}+\frac{2\,\boldsymbol{\lambda} e^{\frac{\mathbf{n}-\mathbf{x}}{2}}}{(\mathbf{1}+e^{\mathbf{n}-\mathbf{x}})}]\right] \\ &\approx \mathbf{x} + \log(\mathbf{1}+e^{\mathbf{n}-\mathbf{x}}) + \frac{\boldsymbol{\lambda}}{\cosh(\frac{\mathbf{n}-\mathbf{x}}{2})},\end{aligned} \tag{4.6}$$

where $\boldsymbol{\lambda}$ is the inner product of the clean speech and noise vectors of mel filter energies in the linear domain, and the last step of approximation uses the assumption that $\boldsymbol{\lambda} \ll \cosh(\frac{\mathbf{n}-\mathbf{x}}{2})$.

In order to simplify the complicated evaluation of the small prediction residual (Eq. (4.6)) of

$$\mathbf{r} = \frac{\boldsymbol{\lambda}}{\cosh(\frac{\mathbf{n}-\mathbf{x}}{2})}, \tag{4.7}$$

an "ignorance" modeling approach is taken to model it as a zero mean, Gaussian random vector. This thus gives a probabilistic model of

$$\mathbf{y} = \mathbf{x} + \mathbf{g}(\mathbf{n} - \mathbf{x}) + \mathbf{r}, \tag{4.8}$$

where $\mathbf{g}(\mathbf{z}) = \log\left(\mathbf{1} + e^{\mathbf{z}}\right)$, and $\mathbf{r} \sim \mathcal{N}(\mathbf{r}; \mathbf{0}, \boldsymbol{\Psi})$.

The Gaussian assumption for the residual $\mathbf{r}$ in Eq. (4.7) allows straightforward computation of the conditional likelihood of the noisy speech vector according to

$$p(\mathbf{y}|\mathbf{x}, \mathbf{n}) = \mathcal{N}[\mathbf{y}; \mathbf{x} + \mathbf{g}(\mathbf{n} - \mathbf{x}), \boldsymbol{\Psi}]. \tag{4.9}$$

We call Eq. (4.9) the Algonquin model, originally developed in [37], where the parameter $\boldsymbol{\Psi}$ is fixed and independent of SNR. This model was later extended to the phase-sensitive model by making $\boldsymbol{\Psi}$ dependent on SNR using more detailed knowledge of the property of speech and noise mixing.

The Algonquin model is a central element of the Algonquin algorithm, which uses prior Gaussian mixture models for both clean speech, $p(\mathbf{x})$, and noise, $p(\mathbf{n})$, in the log domain for computing an MMSE estimate of clean speech $\mathbf{x}$. In deriving the estimate, a variational algorithm is used to obtain an approximate posterior. Multiple point Taylor linearization is used also, one at each of the Gaussian mean vectors in the Gaussian mixture model of clean speech. The Algonquin algorithm in its original form, which uses multiple points of Taylor series expansion, was compared carefully with the single-point expansion method for log-spectral feature enhancement [20,27] in internal evaluation (unpublished). The single-point expansion method requires more iterations to converge but overall is much more efficient than the multiple-point expansion method of Algonquin and does not require the use of variational inference. In terms of performance, the single-point expansion method produces better results, especially after the dynamic prior is introduced to model the temporal correlation in feature enhancement [27].

4.3.2 Step I in Computing Uncertainty: Means

As with the Algonquin model assumption, the following Gaussian-mixture distribution is used as the prior model for clean speech:

$$p(\mathbf{x}_t) = \sum_{m=1}^{M} c_m p(\mathbf{x}_t|m) = \sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_m^x, \boldsymbol{\Sigma}_m^x).$$

For simplicity in presentation without loss of generality, the prior model for noise is assumed to be a time-varying delta function instead of a Gaussian mixture as in the original Algonquin model:

$$p(\mathbf{n}_t) = \delta(\mathbf{n}_t - \bar{\mathbf{n}}_t), \tag{4.10}$$

where $\bar{\mathbf{n}}_t$ is assumed known, and can be determined by any noise tracking algorithm, e.g., [25, 26, 58].

The MMSE estimate is computed as the expected value of the posterior probability $p(\mathbf{x}|\mathbf{y})$:

$$\hat{\mathbf{x}} = E[\mathbf{x}|\mathbf{y}] = \int \mathbf{x} p(\mathbf{x}|\mathbf{y}) d\mathbf{x}. \tag{4.11}$$

Using Bayes' rule and using the prior speech and noise models just described, this MMSE estimate becomes

$$\begin{aligned}\hat{\mathbf{x}} &= \frac{\int \mathbf{x} p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}}{p(\mathbf{y})} \\ &= \frac{\sum_{m=1}^{M} c_m \int\int \mathbf{x} p(\mathbf{n}) p(\mathbf{x}|m) p(\mathbf{y}|\mathbf{x},\mathbf{n}) d\mathbf{x} d\mathbf{n}}{p(\mathbf{y})} \\ &= \frac{\sum_{m=1}^{M} c_m \int \mathbf{x} p(\mathbf{x}|m) p(\mathbf{y}|\mathbf{x},\bar{\mathbf{n}}) d\mathbf{x}}{p(\mathbf{y})}.\end{aligned} \tag{4.12}$$

Substituting the parametric acoustic distortion model of Eq. (4.9) into Eq. (4.12) and carrying out the needed integration in an analytical form via the use of iterative Taylor series approximation (truncation to the first order), we have approximated the evaluation of the MMSE estimate in Eq. (4.12) using the following iterative procedure. First, train and fix all parameters in the clean speech model: $c_m, \boldsymbol{\mu}_m^x$, and $\boldsymbol{\Sigma}_m^x$. Then, compute the noise estimate and the weighting matrices:

$$\begin{aligned}\mathbf{W}_1(m) &= (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi}, \\ \mathbf{W}_2(m) &= \mathbf{I} - \mathbf{W}_1(m).\end{aligned} \tag{4.13}$$

Next, fix the total number, J, of intra-frame iterations. For each frame $t = 2, 3, \ldots, T$ in a noisy utterance $\mathbf{y}_t$, set iteration number $j = 1$, and initialize the clean speech estimate with

$$\hat{\mathbf{x}}_t^{(1)} = \underset{\boldsymbol{\mu}_m^x}{\operatorname{argmax}} \mathcal{N}[\mathbf{y}_t; \boldsymbol{\mu}_m^x + \mathbf{g}(\bar{\mathbf{n}}_t - \boldsymbol{\mu}_m^x), \boldsymbol{\Psi}]. \tag{4.14}$$

Then, execute the following steps for each time frame (and then sequentially over time frames):

- Step 1: Compute

$$\gamma_t^{(j)}(m) = \frac{c_m \mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_m^x + \mathbf{g}^{(j)}, \boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})}{\sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_m^x + \mathbf{g}^{(j)}, \boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})},$$

  where $\mathbf{g}^{(j)} = \log\left(\mathbf{1} + e^{\bar{\mathbf{n}}_t - \hat{\mathbf{x}}_t^{(j)}}\right)$.
- Step 2: Update the MMSE estimate:

$$\hat{\mathbf{x}}_t^{(j+1)} = \sum_m \gamma_t^{(j)}(m) \left[\mathbf{W}_1(m) \boldsymbol{\mu}_m^x + \mathbf{W}_2(m)(\mathbf{y}_t - \mathbf{g}^{(j)})\right]. \tag{4.15}$$

- Step 3: If $j < J$, increment j by 1, and continue the iteration by returning to Step 1. If $j = J$, then increment t by 1 and start the algorithm again by resetting $j = 1$ to process the next time frame until the end of the utterance $t = T$.

The expectation of the enhanced speech feature vector is obtained as the final iteration of the estimate above for each time frame:

$$\boldsymbol{\mu}_{\hat{\mathbf{x}}_t} = \hat{\mathbf{x}}_t^{(J)}. \tag{4.16}$$

4.3.3 Step II in Computing Uncertainty: Variances

Given the expectation for the enhanced speech feature computed as just described, the variance of the enhanced speech feature can now be computed according to

$$\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t} = E[\mathbf{x}_t \mathbf{x}_t^T | \mathbf{y}_t] - \boldsymbol{\mu}_{\hat{\mathbf{x}}_t} \boldsymbol{\mu}_{\hat{\mathbf{x}}_t}^T, \tag{4.17}$$

where the second-order moment is

$$\begin{aligned} E[\mathbf{x}_t \mathbf{x}_t^T | \mathbf{y}_t] &= \int \mathbf{x}_t \mathbf{x}_t^T \, p(\mathbf{x}_t | \mathbf{y}_t, \bar{\mathbf{n}}_t) d\mathbf{x}_t \\ &= \frac{\int \mathbf{x}_t \mathbf{x}_t^T \, p(\mathbf{x}_t) p(\mathbf{y}_t | \mathbf{x}_t, \bar{\mathbf{n}}_t) d\mathbf{x}_t}{p(\mathbf{y}_t)} \\ &= \frac{\sum_{m=1}^{M} c_m \overbrace{\int \mathbf{x}_t \mathbf{x}_t^T \, p(\mathbf{x}_t | m) p(\mathbf{y}_t | \mathbf{x}_t, \bar{\mathbf{n}}_t) d\mathbf{x}_t}^{I_m(\mathbf{y}_t)}}{p(\mathbf{y}_t)}. \end{aligned} \tag{4.18}$$

After using the zero-th order Taylor series to approximate the nonlinear function $\mathbf{g}(\bar{\mathbf{n}}_t - \mathbf{x}_t)$ by $\mathbf{g}_0(\bar{\mathbf{n}}_t - \mathbf{x}_0)$, the integral in Eq. (4.18) becomes

$$\begin{aligned} I_m(\mathbf{y}_t) &\approx \int \mathbf{x}_t \mathbf{x}_t^T \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_m^x, \boldsymbol{\Sigma}_m^x) \mathcal{N}(\mathbf{y}_t; \mathbf{x}_t + \mathbf{g}_0, \boldsymbol{\Psi}) d\mathbf{x}_t \\ &= \int \mathbf{x}_t \mathbf{x}_t^T \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_m^x, \boldsymbol{\Sigma}_m^x) \mathcal{N}(\mathbf{x}_t; \mathbf{y}_t - \mathbf{g}_0, \boldsymbol{\Psi}) d\mathbf{x}_t \\ &= \int \mathbf{x}_t \mathbf{x}_t^T \mathcal{N}\left[\mathbf{x}_t; \boldsymbol{\theta}_m(t), (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})^{-1} \boldsymbol{\Sigma}_m^x \boldsymbol{\Psi}\right] d\mathbf{x}_t N_m(\mathbf{y}_t) \\ &= \left[(\boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})^{-1} \boldsymbol{\Sigma}_m^x \boldsymbol{\Psi} + \boldsymbol{\theta}_m \boldsymbol{\theta}_m^T\right] N_m(\mathbf{y}_t) \end{aligned} \tag{4.19}$$

where

$$\begin{aligned} \boldsymbol{\theta}_m(t) &= (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})^{-1} \left[\boldsymbol{\Psi} \boldsymbol{\mu}_m^x + \boldsymbol{\Sigma}_m^x (\mathbf{y}_t - \mathbf{g}_0)\right], \\ N_m(\mathbf{y}_t) &= \mathcal{N}\left[\mathbf{y}_t - \mathbf{g}_0; \boldsymbol{\mu}_m^x, \boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi}\right] = \mathcal{N}\left[\mathbf{y}_t; \boldsymbol{\mu}_m^x + \mathbf{g}_0, \boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi}\right]. \end{aligned}$$

Substituting the result of Eq. (4.19) into Eq. (4.18), we obtain

$$E[\mathbf{x}_t \mathbf{x}_t^T | \mathbf{y}_t] = \sum_{m=1}^{M} \eta_m(\mathbf{y}_t) \left[(\boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})^{-1} \boldsymbol{\Sigma}_m^x \boldsymbol{\Psi} + \boldsymbol{\theta}_m(t) \boldsymbol{\theta}_m^T(t) \right], \tag{4.20}$$

where

$$\eta_m(\mathbf{y}_t) = \frac{c_m N_m(\mathbf{y}_t)}{\sum_{m=1}^{M} c_m N_m(\mathbf{y}_t)},$$

and where we used the result that $p(\mathbf{y}_t) = \sum_{m=1}^{M} c_m N_m(\mathbf{y}_t)$ for the denominator.

Equation (4.17) then gives the estimate of the variance for the enhanced feature. In the implementation, an iterative procedure is also used to estimate the variance, for the same purpose of reducing errors caused by the approximation of $\mathbf{g}(\bar{\mathbf{n}} - \mathbf{x})$ by $\mathbf{g}_0(\bar{\mathbf{n}} - \mathbf{x}_0)$. For each iteration, the variance estimate takes the final form of

$$\begin{aligned} \boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t} = & \sum_{m=1}^{M} \eta_m(\mathbf{y}_t) \left[(\boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})^{-1} \boldsymbol{\Sigma}_m^x \boldsymbol{\Psi} + \boldsymbol{\theta}_m(t) \boldsymbol{\theta}_m^T(t) \right] \\ & - \left[\sum_m \gamma_t(m) \left(\mathbf{W}_1(m) \boldsymbol{\mu}_m^x + \mathbf{W}_2(m)(\mathbf{y}_t - \mathbf{g}_0) \right) \right] \cdot \\ & \left[\sum_m \gamma_t(m) \left(\mathbf{W}_1(m) \boldsymbol{\mu}_m^x + \mathbf{W}_2(m)(\mathbf{y}_t - \mathbf{g}_0) \right) \right]^T \end{aligned} \tag{4.21}$$

after combining Eqs. (4.17), (4.20), and (4.15). Note that the weights $\gamma_t(m)$ above in the form of posterior probability are computed for each of the iterations.

4.3.4 Discussions

As can be seen from Eqs. (4.21) and (4.16), the first two moments of the probabilistic feature enhancement, characterized by $p(\mathbf{x}|\mathbf{y})$, is dynamic or time-varying on a frame-by-frame basis. This is very powerful, and is difficult to achieve using the common model compensation techniques. To the best of our knowledge, the only other technique that also achieves frame-by-frame compensation with practical success is variable-parameter HMMs [84, 85], but this benefit comes with a higher computation cost.

Characterization of feature uncertainty is typically done using the first two moments of $p(\mathbf{x}|\mathbf{y})$ as shown above. The approach we took as shown above permits the computation of any higher-order moments. But how to incorporate such more detailed information about uncertainty into the decoding rule of speech recognizers based on Gaussian mixture HMMs in a computationally efficient way is an open problem.

One extension of the uncertainty computation technique discussed above is to remove the temporal conditional independence assumption, which has been discussed in some detail in [27, 51]. Two solutions are offered in [27] and [51] both

demonstrating clear performance improvement after incorporating the improved uncertainty estimates.

In [65,67], the authors presented an interesting analysis demonstrating problematic issues with front-end or feature uncertainty decoding schemes, e.g., [5, 19, 32, 50, 51, 57]. The crux of the problem is the often-found acoustic regions where at low SNR no discriminative information is retained since only a single set of compensation parameters is propagated from the front-end processor to the recognizer's decoder. They developed an improved scheme, called joint uncertainty decoding (JUD), where the model-based concept is embedded into uncertainty decoding. Specifically, instead of linking feature components to the recognizer components, they associate each feature component with a set of recognition model components. Introducing the model component sets provides discriminative information at the otherwise non-discriminative acoustic regions in the original front-end uncertainty decoding without use of any discriminative method. How to exploit some concepts from well-established powerful discriminative techniques (e.g., [44]) to further improve joint uncertainty decoding, or simply to improve the more efficient front-end uncertainty decoding, is an interesting research topic.

Another possible extension of the uncertainty computation technique discussed in this section is to use more advanced models of speech distortion than the Algonquin model. A series of such models have been developed, exploiting the SNR dependency of the residual error in the Algonquin model but based on more detailed analysis than the Algonquin model of the phase relationship between clean speech and the mixing noise. We will take up this topic in the next section, with a review on how these phase-sensitive models have been used for fixed-point feature compensation, which has yet to be extended to derive feature compensation uncertainty with these models.

4.4 Use of a Phase-Sensitive Model for Feature Compensation

Traditionally, the interaction model for environmental distortion ignores the phase asynchrony between the clean speech and the mixing noise [1,71]; it is known as the VTS-model. This type of crude model has been improved over the past several years to achieve higher fidelity that removes the earlier simplifying assumption by including random phase asynchrony. As discussed in the preceding section, the Algonquin model is a simple kind of extension that lumps all modeling errors, including the phase effect, into a zero mean, fixed-variance Gaussian residual in an "ignorant" manner [19,27,37,38]. Phase sensitive models developed subsequently make further improvements over the Algonquin model, resulting in an SNR-dependent residual component. A series of work on the phase-sensitive models, including their successful applications to both feature compensation and model compensation, can be found in [21, 23, 31, 61, 62, 75, 81]. In this section, we limit our review on this part of the literature to feature compensation only.

4.4.1 Phase-Sensitive Modeling of Acoustic Distortion — Deterministic Version

To introduce the background and for simplicity, we derive the phase-sensitive model in the log filter bank domain. (This can be easily extended to the cepstral domain.) Using the discrete-time, linear system model for acoustic distortion in the time domain, we have the well-known relationship among noisy speech ($y(t)$), clean speech ($x(t)$), additive noise ($n(t)$), and the impulse response of the linear distortion channel ($h(t)$):

$$y(t) = x(t) * h(t) + n(t).$$

In the frequency domain, the equivalent relationship is

$$Y[k] = X[k]H[k] + N[k], \tag{4.22}$$

where k is the frequency bin index in the DFT given a fixed-length time window, and $H(k)$ is the (frequency domain) transfer function of the linear channel.

The power spectrum of the noisy speech can then be obtained from the DFT in Eq. (4.22) by

$$\begin{aligned} |Y[k]|^2 &= |X[k]H[k] + N[k]|^2 \\ &= |X[k]|^2|H[k]|^2 + |N[k]|^2 + (X[k]H[k])(N[k])^* \\ &\quad + (X[k]H[k])^* N[k] \\ &= |X[k]|^2|H[k]|^2 + |N[k]|^2 + 2|X[k]||H[k]||N[k]|\cos\theta_k, \end{aligned} \tag{4.23}$$

where θ_k denotes the (random) phase angle between the two complex variables $N[k]$ and $(X[k]H[k])$. Equation (4.23) incorporates the phase relationship between the (linearly filtered) clean speech and the additive corrupting noise in the speech distortion process. It is noted that in the traditional, phase-insensitive models for acoustic distortion, the last term in Eq. (4.23) has been assumed to be zero. This is correct only in the expected sense. The phase-sensitive model presented here based on Eq. (4.23) with non-zero instantaneous values in the last term removes this common but unrealistic assumption.

After applying a set of mel scale filters (L in total) to the spectrum $|Y[k]|^2$ in the frequency domain, where the l-th filter is characterized by the transfer function $W_k^{(l)} \geq 0$ (where $\sum_k W_k^{(l)} = 1$), we obtain a total of L mel filter bank energies of

$$\begin{aligned} \sum_k W_k^{(l)}|Y[k]|^2 &= \sum_k W_k^{(l)}|X[k]|^2|H[k]|^2 + \sum_k W_k^{(l)}|N[k]|^2 \\ &\quad + 2\sum_k W_k^{(l)}|X[k]||H[k]||N[k]|\cos\theta_k, \end{aligned} \tag{4.24}$$

with $l = 1, 2, \ldots, L$.

Denoting the various filter bank energies in Eq. (4.24) by

$$|\tilde{Y}^{(l)}|^2 = \sum_k W_k^{(l)} |Y[k]|^2,$$
$$|\tilde{X}^{(l)}|^2 = \sum_k W_k^{(l)} |X[k]|^2,$$
$$|\tilde{N}^{(l)}|^2 = \sum_k W_k^{(l)} |N[k]|^2, \tag{4.25}$$

and

$$|\tilde{H}^{(l)}|^2 = \frac{\sum_k W_k^{(l)} |X[k]|^2 |H[k]|^2}{|\tilde{X}^{(l)}|^2},$$

we simplify Eq. (4.24) to

$$|\tilde{Y}^{(l)}|^2 = |\tilde{X}^{(l)}|^2 |\tilde{H}^{(l)}|^2 + |\tilde{N}^{(l)}|^2 + 2\alpha^{(l)} |\tilde{X}^{(l)}||\tilde{H}^{(l)}||\tilde{N}^{(l)}|, \tag{4.26}$$

where we define the "phase factor" as

$$\alpha^{(l)} \equiv \frac{\sum_k W_k^{(l)} |X[k]||H[k]||N[k]| \cos\theta_k}{|\tilde{X}^{(l)}||\tilde{H}^{(l)}||\tilde{N}^{(l)}|}. \tag{4.27}$$

Since $\cos\theta_k \leq 1$, we have

$$|\alpha^{(l)}| \leq \frac{\sum_k W_k^{(l)} |X[k]||H[k]||N[k]|}{|\tilde{X}^{(l)}||\tilde{H}^{(l)}||\tilde{N}^{(l)}|}.$$

The right-hand side is the normalized inner product of vectors $\bar{N}$ and $\bar{X}^H$, with elements $\bar{N}_k \equiv \sqrt{W_k^{(l)}} |\tilde{N}^{(l)}(k)|$ and $\bar{X}_k^H \equiv \sqrt{W_k^{(l)}} |\tilde{X}^{(l)}(k)||\tilde{H}^{(l)}(k)|$. Hence

$$|\alpha^{(l)}| \leq \frac{<\bar{N}, \bar{X}^H>}{|\bar{N}||\bar{X}^H|} \leq 1.$$

Further, we define the log mel filter bank energy (log spectrum) vectors

$$\mathbf{y} = \begin{bmatrix} \log|\tilde{Y}^{(1)}|^2 \\ \log|\tilde{Y}^{(2)}|^2 \\ \vdots \\ \log|\tilde{Y}^{(l)}|^2 \\ \vdots \\ \log|\tilde{Y}^{(L)}|^2 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} \log|\tilde{X}^{(1)}|^2 \\ \log|\tilde{X}^{(2)}|^2 \\ \vdots \\ \log|\tilde{X}^{(l)}|^2 \\ \vdots \\ \log|\tilde{X}^{(L)}|^2 \end{bmatrix},$$

$$\mathbf{n} = \begin{bmatrix} \log|\tilde{N}^{(1)}|^2 \\ \log|\tilde{N}^{(2)}|^2 \\ \vdots \\ \log|\tilde{N}^{(l)}|^2 \\ \vdots \\ \log|\tilde{N}^{(L)}|^2 \end{bmatrix}, \mathbf{h} = \begin{bmatrix} \log|\tilde{H}^{(1)}|^2 \\ \log|\tilde{H}^{(2)}|^2 \\ \vdots \\ \log|\tilde{H}^{(l)}|^2 \\ \vdots \\ \log|\tilde{H}^{(L)}|^2 \end{bmatrix}, \tag{4.28}$$

and the vector of phase factors

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha^{(1)} \\ \alpha^{(2)} \\ \vdots \\ \alpha^{(l)} \\ \vdots \\ \alpha^{(L)} \end{bmatrix}.$$

Then, we rewrite Eq. (4.26) as

$$\begin{aligned} e^{\mathbf{y}} &= e^{\mathbf{x}} \bullet e^{\mathbf{h}} + e^{\mathbf{n}} + 2\,\boldsymbol{\alpha} \bullet e^{\mathbf{x}/2} \bullet e^{\mathbf{h}/2} \bullet e^{\mathbf{n}/2} \\ &= e^{\mathbf{x}+\mathbf{h}} + e^{\mathbf{n}} + 2\,\boldsymbol{\alpha} \bullet e^{(\mathbf{x}+\mathbf{h}+\mathbf{n})/2}, \end{aligned} \tag{4.29}$$

where the $\bullet$ operation for two vectors denotes element-wise product, and each exponentiation of a vector above is also an element-wise operation. To obtain the log mel filter bank energy for noisy speech, we apply the log operation on both sides of Eq. (4.29):

$$\begin{aligned} \mathbf{y} &= \log\left[e^{\mathbf{x}+\mathbf{h}} \bullet (\mathbf{1} + e^{\mathbf{n}-\mathbf{x}-\mathbf{h}} + 2\boldsymbol{\alpha} \bullet e^{\frac{\mathbf{x}+\mathbf{h}+\mathbf{n}}{2}-\mathbf{x}-\mathbf{h}})\right] \\ &= \mathbf{x} + \mathbf{h} + \log[\mathbf{1} + e^{\mathbf{n}-\mathbf{x}-\mathbf{h}} + 2\boldsymbol{\alpha} \bullet e^{\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2}}] \\ &\equiv \mathbf{y}(\mathbf{x}, \mathbf{n}, \mathbf{h}, \boldsymbol{\alpha}). \end{aligned} \tag{4.30}$$

From Eq. (4.30), the phase factor (vector) $\boldsymbol{\alpha}$ can be solved as a function of the remaining variables:

$$\begin{aligned}\boldsymbol{\alpha} &= \frac{e^{\mathbf{y}-\mathbf{x}-\mathbf{h}} - e^{\mathbf{n}-\mathbf{x}-\mathbf{h}} - \mathbf{1}}{2e^{\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2}}} \\ &= 0.5(e^{\mathbf{y}-\frac{\mathbf{n}+\mathbf{x}+\mathbf{h}}{2}} - e^{\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2}} - e^{-\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2}}) \\ &\equiv \boldsymbol{\alpha}(\mathbf{x},\mathbf{n},\mathbf{h},\mathbf{y}). \end{aligned} \tag{4.31}$$

Equation (4.30) or Eq. (4.31) constitutes the (deterministic) version of the phase-sensitive model for acoustic distortion due to additive noise in the log-spectral domain.

4.4.2 The Phase-Sensitive Model of Acoustic Distortion — Probabilistic Version

We now use the nonlinear relationship between the phase factor $\boldsymbol{\alpha}$ and the log-domain signal quantities of $\mathbf{x},\mathbf{n},\mathbf{h}$, and $\mathbf{y}$, as derived above and shown in Eqs. (4.30) or (4.31), as the basis to develop a probabilistic phase-sensitive model for the acoustic environment. The outcome of a probabilistic model for the acoustic environment is explicit determination of the conditional probability, $p(\mathbf{y}|\mathbf{x},\mathbf{n},\mathbf{h})$, of noisy speech observations ($\mathbf{y}$) given all other variables $\mathbf{x},\mathbf{n}$, and $\mathbf{h}$. This conditional probability is what is required in the Bayesian network model to specify the conditional dependency. This conditional probability is also required for deriving an optimal estimate of clean speech, which was carried out in [23].

To determine the form of $p(\mathbf{y}|\mathbf{x},\mathbf{n},\mathbf{h})$, we first need to assume a form of the statistical distribution for the phase factor $\boldsymbol{\alpha} = \{\alpha^{(l)}, l = 1,2,...,L\}$. To accomplish this, we note that the angle θ_k between the complex variables of $N[k]$ and $(X[k]H[k])$ is uniformly distributed over $(-\pi,\pi)$. This amounts to the maximal degree of randomness in mixing speech and noise, and has been empirically observed to be correct.

Then, from the definition of $\alpha^{(l)}$ in Eq. (4.27), it can be shown that the phase factor $\alpha^{(l)}$ for each mel filter l can be approximated by a (weighted) sum of a number of independent, zero mean random variables $\cos(\theta_k)$ distributed (non-uniformly but symmetrically) over $(-1,1)$, where the total number of terms equals the number of DFT bins (with a non-zero gain) allocated to the mel filter. When the number of terms becomes large, as is typical for high-frequency filters, the central limit theorem postulates that $\alpha^{(l)}$ will be approximately Gaussian. The law of large numbers further postulates that the Gaussian distribution will have a zero mean since each term of $\cos(\theta_k)$ has a zero mean.

Thus, the statistical distribution for the phase factor can be reasonably assumed to be a zero mean Gaussian:

$$p(\alpha^{(l)}) = \mathcal{N}(\alpha^{(l)}; 0, \Sigma_\alpha^{(l)}),$$

where the filter-dependent variance $\Sigma_\alpha^{(l)}$ is estimated from a set of training data. Since noise and (channel-distorted) clean speech are mixed independently for each

DFT bin, we can also reasonably assume that the different components of the phase factor $\boldsymbol{\alpha}$ are uncorrelated. Thus, we have the multivariate Gaussian distribution of

$$p(\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{\alpha}; \mathbf{0}, \boldsymbol{\Sigma}_\alpha), \tag{4.32}$$

where $\boldsymbol{\Sigma}_\alpha$ is a diagonal covariance matrix.

Given $p(\boldsymbol{\alpha})$, we are in a position to derive an appropriate form for $p_y(\mathbf{y}|\mathbf{x},\mathbf{n},\mathbf{h})$. To do so, we first fix the values of $\mathbf{x}, \mathbf{n}$, and $\mathbf{h}$, treating them as constants. We then view Eq. (4.30) as a (monotonic) nonlinear transformation from random variables $\boldsymbol{\alpha}$ to $\mathbf{y}$. Using the well-known result from probability theory on determining the pdf for functions of random variables, we have

$$p_y(\mathbf{y}|\mathbf{x},\mathbf{n},\mathbf{h}) = |J_{\boldsymbol{\alpha}}(\mathbf{y})|\; p_\alpha(\boldsymbol{\alpha}|\mathbf{x},\mathbf{n},\mathbf{h}), \tag{4.33}$$

where $J_{\boldsymbol{\alpha}}(\mathbf{y}) = \frac{1}{\frac{\partial \mathbf{y}}{\partial \boldsymbol{\alpha}}}$ is the Jacobian of the nonlinear transformation.

The diagonal elements of the Jacobian can be computed using Eq. (4.30) and then using Eq. (4.29) by

$$\begin{aligned}
\operatorname{diag}\left(\frac{\partial \mathbf{y}}{\partial \boldsymbol{\alpha}}\right) &= \frac{2e^{\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2}}}{\mathbf{1} + e^{\mathbf{n}-\mathbf{x}-\mathbf{h}} + 2\boldsymbol{\alpha} \bullet e^{\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2}}} \\
&= \frac{2e^{\frac{\mathbf{n}+\mathbf{x}+\mathbf{h}}{2}}}{e^{\mathbf{x}+\mathbf{h}} + e^{\mathbf{n}} + 2\boldsymbol{\alpha} \bullet e^{\frac{\mathbf{n}+\mathbf{x}+\mathbf{h}}{2}}} \\
&= 2\, e^{\frac{\mathbf{n}+\mathbf{x}+\mathbf{h}}{2} - \mathbf{y}}.
\end{aligned} \tag{4.34}$$

The determinant of the diagonal matrix of Eq. (4.34) is then the product of all the diagonal elements.

Also, the Gaussian assumption for $\boldsymbol{\alpha}$ gives

$$p(\boldsymbol{\alpha}|\mathbf{x},\mathbf{n},\mathbf{h}) = p\left[\boldsymbol{\alpha}(\mathbf{x},\mathbf{n},\mathbf{h},\mathbf{y})\right] = \mathcal{N}\left[\boldsymbol{\alpha}(\mathbf{x},\mathbf{n},\mathbf{h},\mathbf{y}); \mathbf{0}, \boldsymbol{\Sigma}_\alpha\right]. \tag{4.35}$$

Substituting Eqs. (4.34) and (4.35) into Eq. (4.33), we establish the following probabilistic model of the acoustic environment:

$$\begin{aligned}
p_y(\mathbf{y}|\mathbf{x},\mathbf{n},\mathbf{h}) = \frac{1}{2}\; &\left| \operatorname{diag}\left(e^{\mathbf{y} - \frac{\mathbf{n}+\mathbf{x}+\mathbf{h}}{2}}\right) \right| \\
&\mathcal{N}\left[\frac{1}{2}(e^{\mathbf{y} - \frac{\mathbf{n}+\mathbf{x}+\mathbf{h}}{2}} - e^{\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2}} - e^{-\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2}}); \mathbf{0}, \boldsymbol{\Sigma}_\alpha\right].
\end{aligned} \tag{4.36}$$

Because $\boldsymbol{\alpha}$ is the inner product (proportional to cosine of the phase) of the mel filter vectors of noise and clean speech characterizing their phase relationship, a Gaussian distribution on it makes the distortion model of Eq. (4.36) phase-sensitive.

4.4.3 Feature Compensation Experiments and Lessons Learned

The phase-sensitive model expressed in the form of Eq. (4.6) is used directly in the MMSE estimate of clean speech, which gives the compensated cepstral features that feed into the speech recognizer. Deng et al. [23] provides details of the MMSE estimate derivation using first-order Taylor series expansion. Second-order Taylor series expansion can also be used to give a somewhat more accurate MMSE estimate without incurring more computation (unpublished).

As reported in [23], a diagnostic experiment was carried out to assess the role of phase asynchrony in feature enhancement for noise-robust speech recognition. To eliminate the factor of noise power estimation inaccuracy, phase-removed true noise power is used since in the Aurora 2 task the true noise's waveforms are made readily available [46]. Table 4.1 lists the percent accuracy in the Aurora 2 standard task of digit recognition (as a function of the feature enhancement algorithm iterations using the phase-sensitive model; see the algorithm in [23]). Clean HMMs (simple back-end) as provided by the Aurora 2 task are used for recognizing enhanced features.

Table 4.1: Percent accurate digit recognition rate for the Aurora 2 task as a function of the feature enhancement algorithm iteration number using the phase-sensitive model. Phase-removed true noise features (noise power spectra) are used in this diagnostic experiment as the n-layer variables

Itrs	1	2	4	7	12
SetA	94.12	96.75	97.96	98.11	98.12
SetB	94.80	97.29	98.10	98.48	98.55
SetC	91.00	94.50	96.50	97.86	98.00
Ave.	93.77	96.52	97.72	98.21	**98.27**

When the phase information is removed, how much does the performance suffer? To examine this issue, several spectral subtraction methods are used where the same phase-removed true noise features are used as in Table 1. After careful tuning of the spectral subtraction parameter of the floor value, the best accuracy is 96% (see detailed results in Table 4.2), significantly below the accuracy of 98% obtained with the use of the phase-sensitive model.

However, when instead of the true noise power, the estimated noise power is used (with the algorithm for noise power estimation described in [23]), improvement of recognition accuracy from the use of the phase-insensitive model to the use of the phase-sensitive model becomes much smaller, from 84.80% to 85.74%; see detailed results in Table 4.3.

What could be the reason for the drastic difference between the performance improvements (from the phase-insensitive to phase-sensitive models) with and without noise estimation errors? Let us examine Eq. (4.26). It is clear that the third, phase-related term and the second, noise-power term are added to contribute to the power

Table 4.2: Performance (percent accurate) for the Aurora 2 task using four versions of spectral subtraction (SS) with the same phase-removed true noise features as in Table 1

Floor	e^{-20}	e^{-10}	e^{-5}	e^{-3}	e^{-2}
SS1	93.57	94.26	**95.90**	92.18	90.00
SS2	12.50	44.00	65.46	88.69	84.44
SS3	88.52	89.26	93.19	90.75	88.00
SS4	10.00	42.50	63.08	87.41	84.26

Table 4.3: Right column: percent accurate digit recognition rates for the Aurora 2 task using noise estimation and phase-sensitive feature enhancement algorithms, both described in [23]. Left column: The baseline results obtained with the phase-insensitive model

	Baseline (no phase)	Enhanced (with phase)
SetA	85.66	86.39
SetB	86.15	86.30
SetC	80.40	83.35
Ave.	**84.80**	**85.74**

of noisy speech. If the estimation error in the second, noise-power term is comparable to the entire third term, then the addition of the third term would not be very meaningful in accounting for the power of noisy speech. This is the most likely explanation for the huge performance improvement when true noise power is used (Tables 4.1 and 4.2) and the relatively mild improvement when noise power estimation contains errors (Table 4.3). The analysis above shows the critical role of noise power estimation in enabling the effectiveness of the phase-sensitive model of environmental distortion.

4.4.4 Discussions

In this section, we described one of the most advanced acoustic distortion models and its application to model-based or structured feature compensation. In terms of the degree of model sophistication, the phase-sensitive model is superior to the Algonquin model, which in turn is superior to the standard deterministic VTS model. For example, as a special case, when the variance from the phase factor is assumed to be independent of SNR, the phase-sensitive model becomes reduced to the Algonquin model. And as the phase factor is eliminated, the model becomes further reduced to the VTS model.

When any of these acoustic distortion models is used for feature enhancement, we call the resulting techniques model-based or structured feature compensation.

This contrasts with the feature normalization techniques widely in use in front-end design for speech recognition, where no such acoustic distortion model is exploited. Examples of the latter include feature moment normalization and cepstral time smoothing. One main difference between these two classes of feature compensation techniques is that the feature normalization methods do not provide any mechanism to determine feature uncertainty. The model-based feature enhancement methods, however, are equipped with this mechanism. In the preceding section, we illustrated how the Algonquin model was used for computing uncertainty in the enhanced features and then for uncertainty decoding. Recently, work has been reported [60] on the use of phase-sensitive models for computing feature uncertainty and then for uncertainty decoding. In this approach, which is detailed in Chapter 8 of this book, the authors propose a phase-sensitive distortion model, where the phase factor α is no longer modeled as a Gaussian and where the moments of α are computed analytically. This distortion model together with an a priori model of clean speech is used to compute the feature posterior by Bayesian inference. The obtained feature posterior is then used for uncertainty decoding.

Finally, we remark that model-based feature compensation also contrasts with the different class of noise-robust speech recognition techniques which we call model-domain compensation, where speech recognition model parameters (e.g., means and variances of the HMM) are modified. We will present model-domain compensation in the next section in the context of studying uncertainty in the model parameter space instead of in the feature space discussed so far.

4.5 Bayesian Decision Rule with Unreliable Model Parameters

4.5.1 Bayesian Predictive Classification Rule

Effective exploitation of uncertainty is a key ingredient in nearly all branches of statistical pattern recognition. In the previous sections, we discussed the uncertainty in the feature domain and examined how the feature-domain uncertainty can be propagated to the back-end speech decoder. In this section, we turn to the more authentic Bayesian framework, which accounts for uncertainty in model parameters, providing a theoretical basis for understanding model compensation.

In the already successful applications of HMM-based speech recognition and speaker verification, uncertainty in the HMM parameter values has been represented by their statistical distributions (e.g., [49, 52, 53]). The motivation of this model-space Bayesian approach has been the widely varied speech properties due to many different sources, including speakers (both intra-speaker and inter-speaker) and acoustic environments, across and possibly within training and test data. In order to take advantage of the model parameter uncertainty, the decision rule for recognition or decoding has been improved from the conventional plug-in MAP rule to the Bayesian predictive classification (BPC) rule [49]:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}} \left[\int_{\Lambda \in \Omega} p(\mathbf{y}|\Lambda, \mathbf{W}) p(\Lambda|\phi, \mathbf{W}) d\Lambda \right] P(\mathbf{W}), \tag{4.37}$$

where ϕ is the set of (deterministic) hyper-parameter characterizing the distribution of the random model parameters, Ω denotes all possible values that the random parameters Λ can take, and the integral becomes the desired acoustic score. How to simplify the integration in Eq. (4.37) to enable robust speech recognition in the model space can be found in [53], which, unfortunately, is not as simple as propagating the uncertainty in the feature space to the effect of only modifying the deterministic HMM variance values as discussed in the preceding sections.

It is possible to combine the effects of model parameter uncertainty and feature uncertainty into a single, comprehensive decision rule. Taking the integrals in both the feature and the model-parameter domains, we have

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}} P(\mathbf{W}) \int_{\Lambda \in \Omega} \int_{\mathbf{x}} \frac{p(\mathbf{x}|\mathbf{W}, \Lambda)}{p(\mathbf{x})} p(\mathbf{x}|\mathbf{y}, \Lambda) p(\Lambda|\phi, \mathbf{x}) d\mathbf{x} d\Lambda. \tag{4.38}$$

4.5.2 Model Compensation Viewed from the Perspective of the BPC Rule

As discussed in [41], there are a large number of model compensation or model-based techniques developed in the past to handle uncertainty in noise-robust speech recognition. It is interesting to place such techniques in the context of the implementation of the BPC rule exemplified in Eq. (4.38).

Let us first simplify Eq. (4.38) by disregarding feature uncertainty; that is, assume the input features have a zero variance value:

$$p(\mathbf{x}|\mathbf{y}, \Lambda) = \delta(\mathbf{y} = \mathbf{x}). \tag{4.39}$$

This then gives

$$\begin{aligned} \hat{\mathbf{W}} &= \underset{\mathbf{W}}{\operatorname{argmax}} P(\mathbf{W}) \int_{\Lambda \in \Omega} \frac{p(\mathbf{y}|\mathbf{W}, \Lambda)}{p(\mathbf{y})} p(\Lambda|\phi, \mathbf{y}) d\Lambda \\ &= \underset{\mathbf{W}}{\operatorname{argmax}} \int_{\Lambda \in \Omega} p(\mathbf{W}|\mathbf{y}, \Lambda) p(\Lambda|\phi, \mathbf{y}) d\Lambda. \end{aligned} \tag{4.40}$$

We further assume that the model-parameter distribution, $p(\Lambda|\phi, \mathbf{y})$, is sharp. It is obvious then that the mode of $p(\Lambda|\phi, \mathbf{y})$ will be at the parameter set which matches best the noisy speech input $\mathbf{y}$. We denote this set of model parameters by $\Lambda(\mathbf{y})$, which is the goal that all the model compensation techniques are searching for. So under the assumption of

$$p(\Lambda|\phi, \mathbf{y}) = \delta(\Lambda = \Lambda(\mathbf{y})), \tag{4.41}$$

we further simplify Eq. (4.40) to

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}}\, p(\mathbf{W}|\mathbf{y}, \Lambda(\mathbf{y})), \tag{4.42}$$

which gives rise to the decision rule of all model compensation techniques published in the literature, where $\Lambda(\mathbf{y})$ is at the mode of the distribution $p(\Lambda|\phi, \mathbf{y})$, or

$$\Lambda(\mathbf{y}) = \underset{\Lambda}{\operatorname{argmax}}\, p(\Lambda|\phi, \mathbf{y}). \tag{4.43}$$

From the above discussion, we can view the decision rule of Eq. (4.40) as a generalized form of model compensation. An implementation of this general form of decision rule and the associated training procedure was carried out to compensate for acoustic variations using generic linear transformations [53]. Structured models for speech and noise interactions such as VTS, Algonquin, or phase-sensitive models have not been explored using the generalized form of model compensation of Eq. (4.40).

Another insight gained from the discussion above is that the generalized form of model compensation (Eq. (4.40)), and further, the generalized form (Eq. (4.38)) that includes both model and feature compensation require active feedback from the speech decoding results to update both speech and noise parameters. Fertile research can be carried out in this direction.

4.6 Model and Feature Compensation — A Taxonomy-Oriented Overview

We now return to the traditional techniques for model-domain compensation where model parameter uncertainty is not considered and where the model parameters are updated as fixed values. We provide an overview aimed at categorizing a large number of existing techniques in model compensation and at clarifying some related nomenclature which sometimes causes confusion in the literature. A similar overview is provided for the counterpart of feature-domain compensation.

4.6.1 Model-Domain Compensation

Many model compensation techniques have been developed by speech recognition researchers over the past 20 years or so. They are used to update the model parameters, e.g., HMM means and variances, contrasting the feature compensation techniques that reduce the noise and other distortions from the speech feature vectors. Model compensation is also called a model-based or model-domain approach, and the goal is to determine from a set of "adaptation" data the model parameter

set $\Lambda(\mathbf{y})$, which has the desirable property of Eq. (4.43), so as to implement the decision rule described by Eq. (4.42).

The techniques of model compensation can be broadly classified into two main categories, depending on two different approaches. In the first category, one uses unstructured, generic transformations to convert the model parameters. The transformations are typically linear, and often, a set of linear transformations are used. The techniques are general, applicable not only to noise compensation but also to other types of acoustic variations, notably speaker adaptation. They involve many parameters and thus require a reasonably large amount of data to estimate them. This category of techniques is also called model adaptation or adaptive scheme, with typical algorithms of

- Maximum Likelihood Linear Regression (MLLR) [40];
- Maximum A Posteriori (MAP) [59];
- constrained MLLR [40];
- noisy constrained MLLR [55]; and
- multi-style training [17, 34].

Multi-style training is classified into this category of unstructured model compensation because there is structured knowledge built into the training, where instead of using a single noise condition, many types of noisy speech are included in the training data. The hope is that one of the types will appear in the deployment condition. Multi-style training updates the model parameters generally in a less efficient and less structured way than does the linear transformation approach. It requires much more training data as a result, and it often produces the updated distributions in HMMs that are unnecessarily broad, making the trained model weakly discriminative.

In the second category of the model compensation techniques, structured transformations are used, which are generally nonlinear and which take into account the way the noisy speech features (e.g., log spectra or cepstra) are produced from the mixing speech and noise. These techniques are sometimes called predictive scheme or structured model adaptation, and the structured transformations are sometimes called mismatch function, interaction model, or acoustic distortion model. In contrast to the unstructured techniques, the structured methods make use of physical knowledge, or a "model", as an approximation to the physics of the speech and noise mixing process. As such, they are not applicable to other types of acoustic variation compensation such as speaker adaptation. Common techniques in the category of structured model compensation include

- Parallel Model Combination (PMC, with log-normal approximation) [39];
- Vector Taylor Series (VTS) [2, 64];
- phase-sensitive model compensation [61–63, 78].

Note that the technique reported in [78] uses linear SPLINE interpolation to approximate the phase-sensitive model presented in Section 4. The division above is based on the type of mismatch or distortion functions in use. Alternative and further divisions can be made based on the kind of approximation exploited to represent the mismatch functions.

One main practical advantage of the structured model compensation techniques over the unstructured counterpart is the much smaller number of free parameters that need to be estimated. However, these parameters, e.g., those in the noise and channel models, are harder to estimate than the generic linear transformation parameters. Typically, second-order approaches are needed to estimate the variance parameters effectively [62, 63]. In addition, the mismatch functions may be inaccurate, and the computational complexity is higher as well.

4.6.2 Feature-Domain Compensation

The same scheme for classifying the model compensation techniques based on structured vs. unstructured methods can be used for feature compensation. The latter is also called feature enhancement or denoising in the literature. The techniques in the category of structured feature compensation involve the use of the same or similar structured transformations as those discussed earlier for model compensation. Because of the use of the distortion or interaction model of speech and noise mixing, structured feature compensation is often called model-based feature enhancement or compensation in the literature. Note that here "model" refers to the distortion model or structured transformation, with examples in Eqs. (4.6), (4.8) and (4.30) ([71, 82]), and it may sometimes also refer to the use of Gaussian mixture models for clean speech. On the other hand, the "model" in model-based compensation refers to the model used for speech recognition, or HMM (e.g., [41]). Some of the commonly used techniques in the category of structured feature-domain compensation have been discussed in Section 3 (the part with mean calculation) and Section 4. Here is a brief summary with some more examples:

- VTS [71];
- Algonquin [20, 37];
- Phase-sensitive modeling for feature enhancement [23, 31, 81].

All these structured enhancement techniques use the estimate of Minimum Mean Squared Error (MMSE) for clean speech in the cepstral or log-spectral domains. In these domains, nonlinearity comes into play. The log domain is believed to be a better one than the linear spectral domain because it is closer to what the back-end HMM speech recognizer receives (see [85] for experimental evidence and related discussions).

In the unstructured category of feature-domain compensation, the techniques developed make no use of structured knowledge of how speech and noise mix expressed in the log domain. They either use stereo data to learn the impact of noise on speech in the log domain implicitly (e.g., SPLICE), or operate in the linear spectral domain, where the mixing process is much simpler (e.g., spectral subtraction). They also include several popular feature normalization methods. Examples of commonly used unstructured feature compensation techniques include

- SPLICE and its extensions, known as stochastic vector mapping [3, 17, 18, 33, 47, 87];
- spectral subtraction [12];
- Wiener filtering combined with the HMM or trajectory-HMM methods [35, 76, 77];
- MMSE estimator on Fourier spectral amplitude [36];
- MMSE estimator on cepstra [85];
- cepstral mean, variance, and histogram normalization (cf., review in [34]); and
- RASTA, FDLP (frequency-domain linear prediction), and other types of modulation spectra (e.g., [72]).

4.6.3 Hybrid Compensation Techniques

We provide a summary in Table 4.4, with entries of classes F1, F2, M1, and M2, of the two-way classification for each of the feature-domain and model-domain compensation techniques discussed so far in this section. The feature and model techniques can be combined to form hybrid techniques, shown as H1 and H2 in Table 4.4.

	Feature Domain	**Model Domain**	**Hybrid**
Un-structured	Class F1	Class M1	Class H1
Structured	Class F2	Class M2	Class H2

Table 4.4: A summary and classification of noise-robust speech recognition techniques

As a summary, some typical examples in each of the classes of techniques in Table 4.4 are provided below again, including the two classes of hybrid techniques to be discussed:

- Class F1: SPLICE, spectral subtraction, Wiener filter, HMM, MMSE, MMSE-Cepstra, CMN (cepstral mean normalization), CVN (cepstral variance normalization), CHN (cepstral histogram normalization), RASTA, modulation spectra;
- Class F2: VTS, Algonquin, phase model;
- Class M1: MLLR, MAP, C-MLLR, N-CMLLR, multi-style training;
- Class M2: PMC, VTS, phase model;
- Class H1: NAT-SS (noise-adaptive training with spectral subtraction), NAT-SPLICE (noise-adaptive training with SPLICE), JAT (or NAT-LR), IVN (irrelevant variability normalization); and
- Class H2: NAT-VTS, UD, JUD.

Examples of structured hybrid techniques are the various uncertainty decoding (UD) techniques discussed in earlier sections in this chapter, where the structured

transformation as provided by the Algonquin model was used to compute the uncertainty in the feature that is subsequently propagated to the HMM decoder. A more comprehensive discussion on uncertainty propagation and decoding is provided in Chapter 3 and other chapters in this book.

The best example of unstructured hybrid techniques is noise-adaptive training (NAT) and its various extensions. Due to its excellent performance and the recent renewed interest in this scheme, we will devote the next section to this important topic.

4.7 Noise Adaptive Training

Noise adaptive training (NAT) is a hybrid strategy of feature compensation and model compensation. The part of feature compensation can be in any form of noise reduction or feature enhancement (structured H2 with F2, or unstructured H1 with F1). The part of model compensation, however, takes the specific form of multi-style (re)training operating on the feature-compensated training data. This original scheme of NAT, first reported in [17], which demonstrated its surprisingly high performance, has formed one of the two standard paradigms (i.e., the multi-style acoustic model of denoised features) in the Aurora experimental framework for the evaluation of speech recognition systems under noisy conditions. Hence, the effectiveness of NAT published in [17] has been verified by at least hundreds of additional experiments with all sorts of feature enhancement techniques and databases worldwide. In addition, the original scheme of NAT has been further developed in various directions during the last decade. In this section, we will provide an overview of these developments.

4.7.1 The Basic NAT Scheme and its Performance

Let us first review and demonstrate the effectiveness of NAT in Figure 4.1 (with data extracted and reorganized from [17]), where the word error rate produced by the author using a standard HMM system (with a noisy 5K-vocabulary Wall Street Journal task) is shown as a function of SNR with added noise. Five sets of results are shown. The results labeled as "Noisy-Noisy matched" (in green) were produced by adding the same noise samples to the training and test data at each SNR level. And the noise used was stationary. In this way, the perfect matching condition was created artificially in the noisy speech domain. This is the condition that perfect model compensation schemes are striving for, and conventional wisdom posits that this sets the upper bound for the system performance. However, when the NAT scheme was developed and applied with two relatively simple unstructured feature compensation techniques, spectral subtraction (SS) and SPLICE, both re-trained HMMs (NAT-SS and NAT-SPLICE) outperform the noisy-matched system under

almost all SNR conditions (except the clean condition with slight degradation). And NAT-SPLICE did better than NAT-SS. This finding has been verified by many types of noise and SNR levels with SPLICE and SS (some of them were published in [17]). Other feature compensation techniques discussed in this chapter were also used with success as a component of the NAT training. More recently, this kind of multi-style training in NAT was also successfully applied to multilingual speech recognition [68].

There have been some discussions among speech recognition researchers as to whether the model or the feature domain is more appropriate for noise compensation. The results discussed above demonstrate that while feature compensation alone may not clearly outperform model compensation (with some possible exception, e.g., [13]), a simple hybrid such as NAT is already sufficient to beat it. The fact that NAT with a more sophisticated type of feature compensation (e.g., SPLICE) outperforms a simpler type (e.g., spectral subtraction), as shown in Figure 4.1, points to the importance of developing higher-quality feature compensation techniques.

In practical deployment, however, the basic paradigm of NAT discussed above is difficult to realize because it requires knowledge of the exact noise characteristics (e.g., noise type and level). Also, if the noise characteristics are numerous, and especially if they are time-varying, NAT cannot be easily carried out in advance even for small systems. More recent developments of the NAT framework have offered possible solutions to this problem, which we will review next.

4.7.2 NAT and Related Work — A Brief Overview

The basic NAT scheme just discussed can be viewed as a way of estimating "clean" speech model parameters, which have been assumed to be available and accurate in all the noise compensation and speech recognition techniques presented so far. However, this assumption rarely holds in practice, since the training data used for building large systems typically contain mixed clean and noisy speech. On the one hand, this situation is similar to speaker adaptive training (SAT) [4], which deals with speaker variation caused by mixed speakers in the training set. On the other hand, the NAT problem differs from SAT in that there is the golden "target" for feature adaptation or compensation, which is truly clean speech. In SAT, there is no such predefined golden "sheep" speaker as the adaptation target.

In the basic NAT scheme of [17], the feature compensation components (SPLICE and SS) have fixed parameters during NAT training. Only the "pseudo-clean" HMM parameters are learned with the objective function, which is independent of SPLICE and SS parameters. And the optimization criterion is maximum likelihood, optimized via EM. This basic scheme has been extended in [48, 54] in two ways. First, the unstructured feature compensation components (SPLICE and SS) are improved to the structured compensation technique of VTS, changing from the unstructured hybrid technique NAT-SPLICE or NAT-SS to the structured hybrid technique NAT-VTS. Second, the free parameters in the feature compensator VTS are subject to

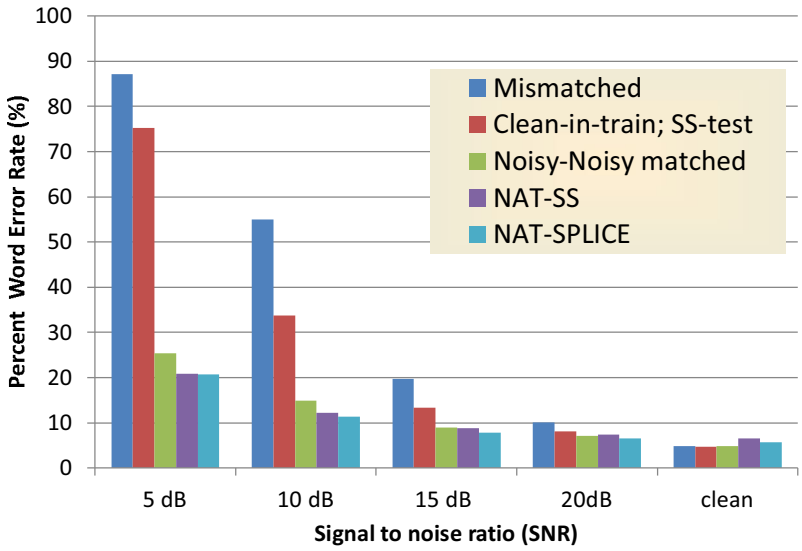

Fig. 4.1: Word error rates for a noisy WSJ task demonstrating that NAT outperforms the best noisy matched condition

joint training with the HMM parameters. These free parameters include the noise means and variances as part of the VTS distortion model. The use of EM for maximum likelihood estimation remains the same. In the work of both [48, 54], the motivation comes from two key insights provided in [17]:

1. Any feature compensation technique inevitably introduces, however small they may be, undesirable residual errors. These errors cause a special kind of model mismatch if no retraining is done; and
2. No absolute clean speech data in training are available, and hence the ultimate models built after the use of any feature compensation technique should settle at best with "pseudo-clean" speech models.

The more recent work of [54] differs from [48] in the way to train the HMM and VTR noise model parameters. One uses iteratively separate steps in the training while insisting on the mapping from noisy speech to "pseudo-clean" speech [48]. The other uses single-step optimization, disregarding the true distribution of the clean speech, matching the best to the adaptation scheme performed at runtime, and achieving somewhat better recognition accuracy on the same task [54]. In this regard, the NAT-VTS version of [54] is very similar to SAT in spirit.

Another interesting extension of NAT is called Joint Adaptation Training (JAT) [66], where linear regression instead of VTS or SPLICE/SS is used to represent the feature transformation. As in [48] and [54], the adaptive transformation in JAT is parameterized, and its parameters are jointly trained with the HMM parameters by the same kind of maximum likelihood criterion as in [17, 48, 54]. Since the feature transformation is linear regression (LR) in the JAT technique, by the nomenclature established in this chapter, JAT can also be rightfully called NAT-LR.

Another set of extension and generalization of the basic NAT scheme is called Irrelevant Variability Normalization (IVN) (e.g., [80,88]), which is a general framework designed not only for noise compensation but also for other types of acoustic mismatch, and it encompasses SAT as a special case. A specific version of IVN [48] designed for noise compensation using VTS as the feature enhancement method, that gives rise to NAT-VTS in our NAT nomenclature, was discussed earlier. In [88], a total of six feature compensation functions are presented, extending the original SPLICE and SS mapping functions proposed in the basic NAT framework [17] in a systematic manner. Among these feature compensation functions are SPLICE, clusters of linear regressions, and clusters of bias removal mappings. And again, joint training of all parameters in the feature compensation functions and in the HMM is carried out in an iterative, two-step fashion. It is noted that in the original NAT-SPLICE of [17], feature compensation is accomplished utterance by utterance in the training set using SPLICE, whose parameters were trained separately from the HMM training for pseudo-clean speech. In the decoding phase, the same SPLICE technique is applied to the test input features. Differently, in IVN of [80,88], the "environment" variable has to be detected, a process called "acoustic sniffing", during the joint HMM and feature compensation parameters' training. The same "acoustic sniffing" is needed in the decoding phase also.

Additional improvements of IVN over the basic NAT framework of [17] are the use of MAP instead of maximum likelihood training [88], and the use of sequential estimation of the feature compensation parameters [80].

Finally, we point out that the mechanism presented in [53] for compensating for extraneous or "irrelevant" variability of spontaneous speech is very similar to that of IVN discussed above. In both techniques, the "condition" or "environment" variable is used to denote a set of discrete unknown or hidden factors underlying the extraneous variations in the observed acoustic data. And they both use a joint training strategy for both the HMM and the transformation parameters. The main difference is that the transformation from the distorted domain to the "canonical" domain is done on the HMM mean vectors in [53] and on the features in IVN.

4.8 Summary and Conclusions

Handling uncertainty in both data and model is a general problem in pattern recognition, and it has special relevance to noise-robust speech recognition, where the uncertainty with both types abounds. In this chapter, selected topics in noise-robust

speech recognition have been presented with in-depth discussion, framed using the Bayesian perspective and centered on the theme of uncertainty treatment.

Since noise robustness is a vast subject, a number of other relevant topics — for example, estimation of noise and channel characteristics, long-term reverberant distortion, fast-changing non-stationary noise tracking, single-microphone speaker separation, and voice activity detection in noisy environments — have not been included in the overview and discussion in this chapter. Also, robust speech recognition performance figures have not been systematically provided in this chapter for comparisons of the techniques presented. The references included in this chapter and other chapters in this book should fill in most of the missing topics, as well as the information about performance comparisons.

Despite significant progress in noise-robust speech recognition over the past two decades or so, many of which have been discussed in this chapter, the problem is far from being solved. Further research in this area is required to enable a sufficiently high level of accuracy in real-world speech recognition applications encompassing a full range of acoustic conditions. Here, a brief discussion is offered from the author's perspective on the expected future research activities in the area of noise-robust speech recognition in the relatively long term.

First, better acoustic modeling than the current ones for speech and noise and the interactions between them is needed. To illustrate this need, we use a simple example here. Estimation of clean speech and mixing noise from noisy speech is a mathematically ill-defined problem, with one known and two unknown variables and one constraint. The only way to have sensible solutions is to impose and exploit prior knowledge as constraints. The hallmark of the Bayesian framework is its formalization of the use of prior knowledge in a mathematically rigorous manner. As elaborated in [9, 10], some powerful sources of prior knowledge in machine speech recognition, especially under noisy environments, come from human speech perception and production. Computational models with appropriate complexity are the first step in exploiting such prior knowledge; e.g., [14–16, 69, 73]. In addition, algorithms extracting key insights from human speech perception and production knowledge and the corresponding computational models are needed to benefit robust speech recognition, e.g., [11, 75]. An example of the benefit can be found in a solution to the multi-talker speech separation and recognition problem, a most difficult kind in robust speech recognition, where the interfering "noise" is another speaker's speech. Powerful graphical modeling and related algorithms, coupled with the Algonquin and phase-sensitive interaction models (see Sections 4.3.1 and 4.4), as well as the use of speech and noise dynamics as part of prior knowledge built into the Bayesian network, offer recognition accuracy superior to that of humans [45]. Given that embedding only very crude dynamic characteristics of speech and noise gives promising results, exploitation of more insightful and relevant knowledge is expected to be fruitful research.

Second, related to the above discussion on better acoustic modeling for speech and noise interactions, we need to develop better techniques to characterize and exploit the effects of noise on speech in a wide variety of front-end features. All the work reviewed in this chapter assumes the use of log spectra or cepstra (e.g.,

MFCC), based on which all nonlinear structured acoustic distortion models are developed and approximated. However, state-of-the-art front-end features may not always be strict cepstra. Tandem features [72], PLP features, some discriminatively derived features [74], and mean-normalized MFCCs often perform better than plain MFCCs. But these features make it difficult to derive the structured distortion models to enable structured feature and model compensation (classes of F2 and M2 in Table 4.4 as well as NAT-VTR). Further, an emerging technology developed from machine learning, deep learning (e.g., [42]), which has just started entering the field of speech recognition [29,70,86], proposes a fundamentally different way of feature extraction from the largely handcrafted features such as MFCC. The layer-by-layer feature extraction strategy in deep learning provides the opportunity to automatically derive powerful features from the primitive raw data, e.g., speech waveform or Fourier transform. Given that in the linear domains of waveform and linear spectrum speech and noise interaction models become much simpler than in MFCC, but that there are some special difficulties involved in using the primitive feature of waveform in speech recognition (e.g., [79]), how to do noise robust speech recognition in the deep learning framework will require complete rethinking of the HMM framework, which we have been so familiar with as the implicit assumption in this and many other chapters in this book.

Third, as elaborated in Section 4.7, the hybrid strategy of feature and model compensation exemplified by NAT, JAT, and IVN is a powerful framework that can handle not only noise-induced uncertainty but also other types of variations extraneous to phonetic distinction. Better integrated algorithm design, improved joint optimization of model and transformation parameters, more effective tracking techniques for time-varying distortion factors, and clever use of metadata as labels for the otherwise hidden "distortion condition" variables that are frequently available from specific speech applications are all fruitful research directions.

References

1. A. Acero: Acoustical and Environmental Robustness in Automatic Speech Recognition. Kluwer Academic Publishers (1993)
2. A. Acero, L. Deng, T. Kristjansson, and J. Zhang: HMM adaptation using vector Taylor series for noisy speech recognition. In: Proc. ICSLP, vol.3, pp. 869-872 (2000)
3. M. Afify, X. Cui, and Y. Gao: Stereo-based stochastic mapping for robust speech recognition. In: Proc. ICASSP (2007)
4. T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul: A compact model for speaker-adaptive training. In: Proc. ICSLP (1996)
5. J. Arrowood and M. Clements: Using observation uncertainty in HMM decoding. In: Proc. ICSLP, Denver, Colorado (2002)
6. R. F. Astudillo, D. Kolossa, and R. Orglmeister: Accounting for the uncertainty of speech estimates in the complex domain for minimum mean squared error speech enhancement. In: Proc. Interspeech (2009)
7. H. Attias, Li Deng, Alex Acero, and John Platt: A new method for speech denoising and robust speech recognition using probabilistic models for clean speech and for noise. In: Proc. of the Eurospeech Conference (2001)

8. H. Attias, J. Platt, Alex Acero, and Li Deng: Speech denoising and dereverberation using probabilistic models. In: Proc. NIPS (2000)
9. J. Baker, Li Deng, Jim Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy: Research developments and directions in speech recognition and understanding. IEEE Signal Processing Magazine, vol. 26, no. 3, pp. 75-80 (2009)
10. J. Baker, Li Deng, S. Khudanpur, C.-H. Lee, J. Glass, N. Morgan, and D. O'Shaughnessy: Updated MINDS report on speech recognition and understanding. IEEE Signal Processing Magazine, vol. 26, no. 4 (2009)
11. J. Bilmes and C. Bartels: Graphical model architectures for speech recognition. IEEE Signal Processing Magazine, vol. 22, no. 5, pp. 89-100 (2005)
12. S.F. Boll: Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. on Acoustics, Speech, and Signal Processing, 27:113-120 (1979)
13. K. Demuynck, X. Zhang, D. Van Compernolle, and H. Van hamme: Feature versus model based noise robustness. In: Proc. Interspeech (2010)
14. L. Deng: Computational models for auditory speech processing. In: Computational Models of Speech Pattern Processing, (NATO ASI Series), pp. 67-77, Springer Verlag (1999)
15. L. Deng: Computational models for speech production. Computational Models of Speech Pattern Processing, (NATO ASI Series), pp. 199-213, Springer Verlag (1999)
16. L. Deng, D. Yu, and A. Acero: Structured speech modeling. IEEE Trans. on Audio, Speech and Language Processing (Special Issue on Rich Transcription), vol. 14, No. 5, pp. 1492-1504 (2006)
17. L. Deng, A. Acero, M. Plumpe, and X.D. Huang: Large vocabulary speech recognition under adverse acoustic environments. In: Proc. ICSLP, pp. 806-809 (2000)
18. L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang: High-performance robust speech recognition using stereo training data. In: Proc. ICASSP, Salt Lake City, Utah (2001)
19. L. Deng, J. Droppo, and A. Acero: Exploiting variances in robust feature extraction based on a parametric model of speech distortion. In: Proc. ICSLP (2002)
20. Li Deng, Jasha Droppo, and Alex Acero: A Bayesian approach to speech feature enhancement using the dynamic cepstral prior. In: Proc. ICASSP, Orlando, Florida (2002)
21. L. Deng, J. Droppo, and A. Acero: Log-domain speech feature enhancement using sequential MAP noise estimation and a phase-sensitive model of the acoustic environment. In: Proc. ICSLP, Denver, Colorado (2002)
22. L. Deng, K. Wang, A. Acero, H. Hon, J. Droppo, C. Boulis, Y. Wang, D. Jacoby, M. Mahajan, C. Chelba, and XD. Huang: Distributed speech processing in MiPad's multimodal user interface. IEEE Trans. on Speech and Audio Processing, vol. 10, no. 8, pp. 605-619 (2002)
23. L. Deng, J. Droppo, and A. Acero: Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise. IEEE Trans. on Speech and Audio Processing, vol.12, no. 2, pp. 133-143 2004)
24. Li Deng and Xuedong Huang: Challenges in adopting speech recognition. Communications of the ACM, vol. 47, no. 1, pp. 11-13, (2004)
25. Li Deng, Jasha Droppo, and Alex Acero: Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition. IEEE Trans. on Speech and Audio Processing, vol. 11, no. 6, pp. 568-580 (2003)
26. Li Deng, Jasha Droppo, and Alex Acero: Incremental Bayes Learning with Prior Evolution for Tracking Non-Stationary Noise Statistics from Noisy Speech Data. In: Proc. ICASSP, Hong Kong (2003)
27. Li Deng, Jasha Droppo, and Alex Acero: Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features. IEEE Trans. on Speech and Audio Processing, vol. 12, no. 3, pp. 218-233 (2004)
28. L. Deng, J. Droppo, and A. Acero: Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. IEEE Trans. on Speech and Audio Processing, vol. 12, no. 3, (2005)
29. Li Deng, Mike Seltzer, Dong Yu, Alex Acero, A. Mohamed, and Geoff Hinton: Binary coding of speech spectrograms using a deep auto-encoder. In: Proc. Interspeech (2010)

30. J. Droppo, A. Acero, and L. Deng: Efficient online acoustic environment estimation for FCDCN in a continuous speech recognition system. In: Proc. ICASSP, Salt Lake City, Utah (2001)
31. J. Droppo, A. Acero, and L. Deng: A nonlinear observation model for removing noise from corrupted speech log Mel-spectral energies. In: Proc. ICSLP, Denver, Colorado (2002)
32. J. Droppo, A. Acero, and L. Deng: Uncertainty decoding with SPLICE for noise robust speech recognition. In: Proc. ICASSP, Orlando, Florida (2002)
33. J. Droppo, L. Deng, and A. Acero: Evaluation of SPLICE on the Aurora 2 and 3 Tasks. In: Proc. ICSLP, Denver, Colorado (2002)
34. J. Droppo and A. Acero: Environmental Robustness. In: Handbook of Speech Processing, Springer (2007)
35. Y. Ephraim: A Bayesian estimation approach for speech enhancement using hidden Markov models. IEEE Trans. on Acoustics, Speech, and Signal Processing, 40:725-735 (1992)
36. Y. Ephraim and D. Malah: Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-32, no. 6, pp. 1109-1121 (1984)
37. B. Frey, L. Deng, A. Acero, and T.T. Kristjansson: Algonquin: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition. In: Proc. Eurospeech, Aalborg, Denmark (2001)
38. B. Frey, T. Kristjansson, Li Deng, and Alex Acero: Learning dynamic noise models from noisy speech for robust speech recognition. In: Proc. Advances in Neural Information Processing Systems (NIPS), vol. 14, Vancouver, Canada, 2001, pp. 101-108 (2001)
39. M.J.F. Gales and S.J. Young: Robust speech recognition in additive and convolutional noise using parallel model combination. Computer Speech and Language, 9:289-307 (1995)
40. M. J. F. Gales: Maximum Likelihood Linear Transformations For HMM-Based Speech Recognition. Computer Speech and Language, 12 (January 1998)
41. M.J.F. Gales: Model-based approaches to handling uncertainty. Chapter 5 of this book (2011)
42. G. Hinton, S. Osindero, and Y. Teh: A fast learning algorithm for deep belief nets. Neural Computation, vol. 18, pp. 1527-1554, 2006)
43. R. Haeb-Umbach and V. Ion: Soft features for improved distributed speech recognition over wireless networks. In: Proc. Interspeech (2004)
44. X. He, L. Deng, and W. Chou: Discriminative learning in sequential pattern recognition — A unifying review. IEEE Signal Processing Magazine (2008)
45. J. Hershey, S. Rennie, P. Olsen, and T. Kristjansson: Super-human multi-talker speech recognition: A graphical modeling approach. Computer Speech and Language (June 2010)
46. H. G. Hirsch and D. Pearce: The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proc. ISCA ITRW ASR (2000)
47. C. Hsieh and C. Wu: Stochastic vector mapping-based feature enhancement using prior-models and model adaptation for noisy speech recognition. Speech Communication, vol. 50, No. 6, pp. 467-475 (2008)
48. Y. Hu and Q. Huo: Irrelevant variability normalization based HMM training using VTS approximation of an explicit model of environmental distortions. In: Proc. Interspeech (2007)
49. C.-H. Lee and Q. Huo: On adaptive decision rules and decision parameter adaptation for automatic speech recognition. Proc. of the IEEE, vol. 88, No. 8, pp. 1241-1269 (2000)
50. V. Ion and R. Haeb-Umbach: Uncertainty decoding for distributed speech recognition over error-prone networks. Speech Communication, vol. 48, pp. 1435-1446 (2006)
51. V. Ion and R. Haeb-Umbach: A novel uncertainty decoding rule with applications to transmission error robust speech recognition. IEEE Trans. Speech and Audio Processing, vol. 16. No. 5, pp. 1047-1060 (2008)
52. H. Jiang and Li Deng: A Bayesian approach to the verification problem: Applications to speaker verification. IEEE Trans. Speech and Audio Proc., vol. 9, No. 8, pp. 874-884 (2001)
53. H. Jiang and L. Deng: A robust compensation strategy against extraneous acoustic variations in spontaneous speech recognition. IEEE Trans. on Speech and Audio Processing, vol. 10, no. 1, pp. 9-17 (2002)

54. O. Kalinli, M.L. Seltzer, and A. Acero: Noise adaptive training using a vector Taylor series approach for noise robust automatic speech recognition. In: Proc. ICASSP, pages 3825-3828, Taipei, Taiwan (2009)
55. D. Kim and M. Gales: Noisy constrained maximum likelihood linear regression for noise robust speech recognition. IEEE Trans. Audio Speech and Language Processing (2010)
56. D.Y. Kim, C.K. Un, and N.S. Kim: Speech recognition in noisy environments using first-order vector Taylor series. Speech Communication, vol. 24, pp. 39-49 (1998)
57. T.T. Kristjansson and B.J. Frey: Accounting for uncertainty in observations: A new paradigm for robust speech recognition. In: Proc. ICASSP, Orlando, Florida (2002)
58. T.T. Kristjansson, B. Frey, L. Deng, and A. Acero: Towards non-stationary model-based noise adaptation for large vocabulary speech recognition. In: Proc. ICASSP (2001)
59. C.-H. Lee: On stochastic feature and model compensation approaches to robust speech recognition. Speech Communication, vol. 25, pp. 29-47 (1998).
60. V. Leutnant and R. Haeb-Umbach: An analytic derivation of a phase-sensitive observation model for noise robust speech recognition. In: Proc. Interspeech (2009)
61. J. Li, D. Yu, Y. Gong, and Li Deng: Unscented Transform with Online Distortion Estimation for HMM Adaptation. In: Proc. Interspeech (2010)
62. J. Li, D. Yu, L. Deng, Y. Gong, and A. Acero: A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions. Computer Speech and Language, vol. 23, pp. 389-405 (2009)
63. J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero: HMM Adaptation Using a Phase-Sensitive Acoustic Distortion Model for Environment-Robust Speech Recognition. In: Proc. ICASSP, Las Vegas (2008)
64. J. Li, L. Deng, D. Yu, J. Wu, Y. Gong, and A. Acero: Adaptation of compressed HMM parameters for resource-constrained speech recognition. In: Proc. ICASSP, Las Vegas (2008)
65. H. Liao and M. J. F. Gales: Issues with uncertainty decoding for noise robust speech recognition. In: Proc. ICSLP, pp. 1121-1124 (2006)
66. H. Liao and M. J. F. Gales: Adaptive training with joint uncertainty decoding for robust recognition of noisy data. In: Proc. ICASSP, vol. IV, pp. 389-392 (2007)
67. H. Liao and M.J.F. Gales: Joint uncertainty decoding for noise robust speech recognition. In: Proc. Interspeech (2005)
68. Hui Lin, Li Deng, Dong Yu, Yifan Gong, Alex Acero, and Chi-Hui Lee: A study on multilingual acoustic modeling for large vocabulary ASR. In: Proc. ICASSP (2009)
69. R. Lyon: Machine hearing: An emerging field. IEEE Signal Processing Magazine (September 2010)
70. A. Mohamed, D. Yu, and L. Deng: Investigation of full-sequence training of deep belief networks for speech recognition. In: Proc. Interspeech (2010)
71. P. Moreno: Speech Recognition in Noisy Environments. Ph.D. Thesis, Carnegie Mellon University (1996)
72. N. Morgan et al.: Pushing the envelope — Aside. IEEE Signal Processing Magazine, vol. 22, No. 5, pp. 81-88 (2005)
73. R. Munkong and B.-H. Juang: Auditory perception and cognition — Modularization and integration of signal processing from ears to brain. IEEE Signal Processing Magazine, vol. 25, No. 3, pp. 98-117 (2008)
74. C. Rathinavalu and L. Deng: HMM-based speech recognition using state-dependent, discriminatively derived transforms on Mel-warped DFT features. IEEE Trans. on Speech and Audio Processing, pp. 243-256 (1997)
75. S. Rennie, J. Hershey, P. Olsen: Combining variational methods and loopy belief propagation for multi-talker speech recognition. IEEE Signal Processing Magazine, Special issue of Graphical Models for Signal Processing (Eds. M. Jordan et al.), (November 2010)
76. H. Sameti, H. Sheikhzadeh, Li Deng, and R. Brennan: HMM-based strategies for enhancement of speech signals embedded in nonstationary noise. IEEE Trans. on Speech and Audio Processing, vol. 6, no. 5, pp. 445-455 (1998)
77. H. Sameti and Li Deng: Nonstationary-state hidden Markov model representation of speech signals for speech enhancement. Signal Processing, vol. 82, pp. 205-227 (2002)

78. M. Seltzer, K. Kalgaonkar, and A. Acero: Acoustic model adaptation via linear spline interpolation for robust speech recognition. In: Proc. ICASSP (2010)
79. H. Sheikhzadeh and Li Deng: Waveform-based speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization. IEEE Trans. on Speech and Audio Processing, vol. 2, no. 1, pp. 80-91 (1994)
80. G. Shi, Y. Shi, and Q. Huo: A study of irrelevant variability normalizataion based training and unsupervised online adaptation for LVCSR. In: Proc. Interspeech, Makuhari, Japan (2010)
81. V. Stouten,, H. Van hamme, P. Wambacq: Effect of phase-sensitive environment model and higher order VTS on noisy speech feature enhancement. In: Proc. ICASSP, pp. 433- 436 (2005)
82. V. Stouten, H. Van hamme, and P. Wambacq: Accounting for the uncertainty of speech estimates in the context of model-based feature enhancement. In: Proc. ICSLP, pp. 105-108, Jeju Island, Korea (2004)
83. D. Yu, Li Deng, Yifan Gong, and Alex Acero: A novel framework and training algorithm for variable-parameter hidden Markov models. IEEE Trans. on Audio, Speech and Language Processing, vol. 17, no. 7, pp. 1348-1360, IEEE (2009)
84. D. Yu and Li Deng: Solving nonlinear estimation problems using Splines. IEEE Signal Processing Magazine, vol. 26, no. 4, pp. 86-90, (2009)
85. D. Yu, Li Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero: Robust speech recognition using cepstral minimum-mean-square-error noise suppressor. IEEE Trans. Audio, Speech, and Language Processing, vol. 16, no. 5 (2008)
86. D. Yu and L. Deng: Deep-Structured Hidden Conditional Random Fields for Phonetic Recognition. In: Proc. Interspeech (2010)
87. D. Zhu and Q. Huo: A maximum likelihood approach to unsupervised online adaptation of stochastic vector mapping function for robust speech recognition. In: Proc. ICASSP (2007)
88. D. Zhu and Q. Huo: Irrelevant variability normalization based HMM training using MAP estimation of feature transforms for robust speech recognition. In: Proc. ICASSP (2008)