MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Factorial Models for Noise Robust Speech Recognition

Hershey, J.R.; Rennie, S.J.; Le Roux, J.

TR2012-002 January 2012

Abstract

Noise compensation techniques for robust automatic speech recognition (ASR) attempt to improve system performance in the presence of interference from acoustic signals in the environment other than the speech being recognized. In feature-based noise compensation, which includes speech enhancement, the features extracted from the noisy speech signal are modified before being sent to the recognizer by attempting to remove the effects of noise on the speech features. These methods are discussed in Chapter 12. Model compensation approaches, in contrast, are concerned with extending the acoustic model of speech to account for the effects of noise. A taxonomy of different approaches to noise compensation is depicted in Figure 1.1, which serves as a road map to the present discussion.

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Copyright © Mitsubishi Electric Research Laboratories, Inc., 2012
201 Broadway, Cambridge, Massachusetts 02139

12

Factorial Models for Noise Robust Speech Recognition

John R. Hershey[1], Steven J. Rennie[2], Jonathan Le Roux[1]

[1]*Mitsubishi Electric Research Laboratories*
[2]*IBM Thomas J. Watson Research Center*

12.1 Introduction

Noise compensation techniques for robust automatic speech recognition (ASR) attempt to improve system performance in the presence of acoustic interference. In *feature-based* noise compensation, which includes *speech enhancement* approaches, the acoustic features that are sent to the recognizer are first processed to remove the effects of noise (see Chapter 9). *Model compensation* approaches, in contrast, are concerned with modifying and even extending the acoustic model of speech to account for the effects of noise. A taxonomy of the different approaches to noise compensation is depicted in Figure 12.1, which serves as a road map for the present discussion.

The two main strategies used for model compensation approaches are *model adaptation* and *model-based noise compensation*. Model adaptation approaches implicitly account for noise by adjusting the parameters of the acoustic model of speech, whereas model-based noise compensation approaches explicitly model the noise and its effect on the noisy speech features. Common adaptation approaches include *maximum likelihood linear regression* (MLLR) [56], *maximum a posteriori* (MAP) adaptation [32], and their generalizations [17, 29, 47]. These approaches, which are discussed in Chapter 11, alter the speech acoustic model in a completely data-driven way given additional training data or test data. Adaptation methods are somewhat more general than model-based approaches in that they may handle effects on the signal that are difficult to explicitly model, such as nonlinear distortion and changes in the voice in reaction to noise (the Lombard effect [53]). However, in the presence of additive noise, failing to take take into account the known interactions between speech and noise can be detrimental to performance.

Model-based noise compensation approaches, in contrast to adaptation approaches, explicitly model the different factors present in the acoustic environment: the speech, the various sources of acoustic interference, and how they interact to form the noisy speech

Techniques for Noise Robustness in Automatic Speech Recognition Tuomas Virtanen, Rita Singh, and Bhiksha Raj, Eds.
© 2012 John Wiley & Sons, Ltd

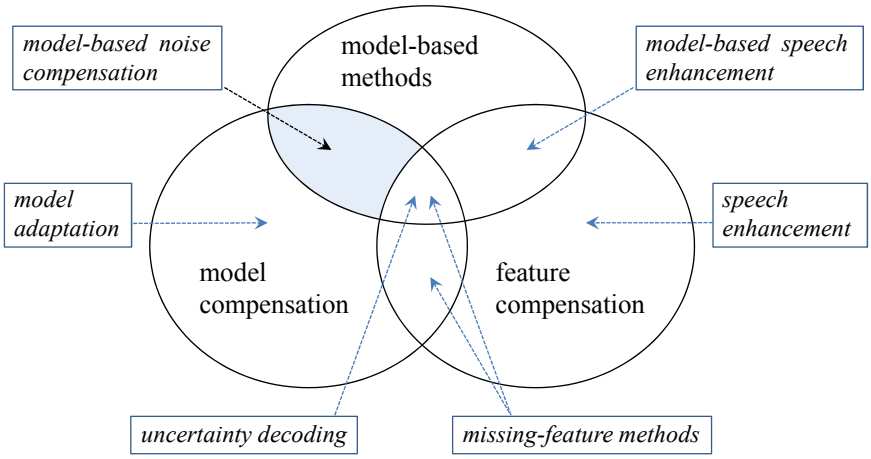


Figure 12.1 Noise compensation methods in a Venn diagram. The shaded region represents model-based noise compensation, the subject of this chapter. Note that the term "model" in "model compensation" refers to the recognizer's acoustic model, whereas in "model-based noise compensation," it refers to the models of additive noise.

signal. By modeling the noise separately from the speech, these factorial models can generalize to combinations of speech and noise sounds not seen during training, and can explicitly represent the dynamics of individual noise sources. A significant advantage of this approach is that the compensated speech features and recognition result are jointly inferred, unlike in feature-based approaches. The recognizer's model of speech dynamics can be directly employed to better infer the acoustic states and parameters of the interference model. Similarly, the model of acoustic interference and its dynamics can be utilized to more accurately estimate the sequence of states of the speech model. Performing these inference processes jointly allows the recognizer to consider different possible combinations of speech and interference.

Approaches that lie somewhere between feature-based and model-based noise compensation include uncertainty decoding [19, 57], which is discussed in Chapter 17, and missing-feature methods [85, 68], which are discussed in Chapters 14, 15 and 16. These methods involve additional communication from the feature enhancement algorithm to the recognizer about the uncertainty associated with the enhanced features being estimated. Model-based compensation approaches can be seen as taking the idea of uncertainty decoding to its logical conclusion: by placing the enhancement model inside the recognizer, the information about uncertainty is considered jointly in terms of the noise model and the full speech model of the recognizer.

Difficult obstacles must be overcome in order to realize the full benefit of the model-based approach. A primary challenge is the complexity of inference: if implemented naively, joint inference in factorial models requires performing computations for each combination of the states of the models. Because of the potential combinatorial explosion, this is prohibitively expensive for many real applications. Alleviating these problems continues to be a core

challenge in the field, and therefore efficient inference is a central theme in this chapter.

Another challenge is the dilemma of feature domains. In feature domains where the interaction between speech and noise is additive, isolating the phonetic content of the speech signal can be difficult. This is because phonetic content is imparted to speech by the filtering effect of the vocal tract, which is approximately multiplicative in the power spectrum. However, in the log spectrum domain the vocal tract filter is additive. Speech recognizers exploit this by using features that are linear transforms of the log spectrum domain. In such domains, the effect of noise is nonlinear, and compensating for it becomes difficult. As such, a major focus of research has been to derive tractable inference algorithms by approximating the interaction between speech and noise in the log spectrum domain.

This chapter presents the fundamental concepts and current state of the art in model-based compensation, while hinting along the way at potential future directions. [1] First, the general framework of the model-based approach is introduced. This is followed by a review of the feature domains commonly used for representing signals, focusing on the way in which additive signals interact deterministically in each domain. A probabilistic perspective on these interaction functions and their approximations is then presented. Following this, several commonly used inference methods which utilize these approximate interaction functions are described in detail. Because computational complexity is of paramount importance in speech processing, we also describe an array of methods which can be used to alleviate the complexity of evaluating factorial models of noisy speech. The chapter concludes with a discussion of many promising research directions in this exciting and rapidly evolving area, with a focus on how complex and highly structured models of noise can be utilized for robust speech recognition.

12.2 The Model-Based Approach

Model-based approaches start with probabilistic models of the features of speech and the noise, and combine them using an *interaction model*, which describes the distribution of the observed noisy speech given the speech and noise. To make this explicit we will need some notation: $p(x)$ denotes a probability distribution. In the case that x is a discrete random variable, p denotes a *probability mass function*, and if x is a continuous random variable, it denotes a *probability density function* (pdf). To simplify notation, we shall specify the random variable considered as a subscript, for example, $p_{\mathsf{x}}(x)$, only when required to avoid confusion. Assume that we have probabilistic models for the features of the clean speech, $\boldsymbol{x}_t$, and the noise $\boldsymbol{n}_t$ at time t: $p(\boldsymbol{x}_t|s_t^{\mathsf{x}})$ and $p(\boldsymbol{n}_t|s_t^{\mathsf{n}})$, which depend on some states s_t^{x} and s_t^{n}. In the context of speech recognition, the clean speech model is typically a hidden Markov model (HMM), which describes the dynamical properties of speech via transition probabilities over the unobserved states s_t^{x}. The interaction model then describes the conditional probability of the noisy speech given the clean speech and the noise, $p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{n}_t)$. Inference in the model-based method involves computing one or more of the following basic quantities: the *state likelihood* $p(\boldsymbol{y}_t|s_t^{\mathsf{x}}, s_t^{\mathsf{n}})$, the *joint clean speech and noise posterior* $p(\boldsymbol{x}_t, \boldsymbol{n}_t|\boldsymbol{y}_t, s_t^{\mathsf{x}}, s_t^{\mathsf{n}})$, and the *clean speech estimate* $\mathrm{E}(\boldsymbol{x}_t|\boldsymbol{y}_t, s_t^{\mathsf{x}}, s_t^{\mathsf{n}})$ for a given hypothesis of the speech and noise states s_t^{x} and s_t^{n}. The state likelihood, which is needed in speech recognition to compute the

[1] Additional perspectives and background material may be found in recent reviews on this topic [13, 30].

posterior probability of state sequences, involves the integral

$$p(\boldsymbol{y}_t|s_t^{\mathsf{x}}, s_t^{\mathsf{n}}) = \int p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{n}_t)p(\boldsymbol{n}_t|s_t^{\mathsf{n}})p(\boldsymbol{x}_t|s_t^{\mathsf{x}})\ \mathrm{d}\boldsymbol{x}_t\ \mathrm{d}\boldsymbol{n}_t. \tag{12.1}$$

The joint posterior of the speech and noise features can be computed using the above integral:

$$p(\boldsymbol{x}_t, \boldsymbol{n}_t|\boldsymbol{y}_t, s_t^{\mathsf{x}}, s_t^{\mathsf{n}}) = \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{n}_t)p(\boldsymbol{n}_t|s_t^{\mathsf{n}})p(\boldsymbol{x}_t|s_t^{\mathsf{x}})}{p(\boldsymbol{y}_t|s_t^{\mathsf{x}}, s_t^{\mathsf{n}})}. \tag{12.2}$$

The expected value of the speech features, used in feature-based compensation, can then be obtained as follows:

$$\mathrm{E}(\boldsymbol{x}_t|s_t^{\mathsf{x}}, s_t^{\mathsf{n}}) = \int \boldsymbol{x}_t\ p(\boldsymbol{x}_t, \boldsymbol{n}_t|\boldsymbol{y}_t, s_t^{\mathsf{x}}, s_t^{\mathsf{n}})\ \mathrm{d}\boldsymbol{x}_t\ \mathrm{d}\boldsymbol{n}_t. \tag{12.3}$$

For uncertainty-decoding approaches, a measure of uncertainty such as the posterior variance, $\mathrm{Var}(\boldsymbol{x}_t|s_t^{\mathsf{x}}, s_t^{\mathsf{n}})$, would also need to be computed (see Chapter 17 for more details). Note that there are typically mixture components for each state, so that $p(\boldsymbol{x}_t|s_t^{\mathsf{x}}) = \sum_{c_t^{\mathsf{x}}} p(\boldsymbol{x}_t|c_t^{\mathsf{x}})p(c_t^{\mathsf{x}}|s_t^{\mathsf{x}})$. In the rest of this chapter, we neglect mixture components to avoid clutter, as introducing them is straightforward and irrelevant to the main problem of computing the above integrals.

Given this general framework, what remains is to show how the above integrals can be accurately and efficiently estimated in the feature domains commonly used in speech modeling. To that end, we turn to the interaction functions that result from analysis of signals in different feature domains.

12.3 Signal Feature Domains

We shall present here the different representations of a signal commonly involved in automatic speech recognition, introduce the corresponding notations, and describe the interaction functions between clean speech and noise in each domain. Due to the complexity of these interactions, and in particular due to the nonlinear transformations involved, approximations are often required. We shall point them out as we proceed, and mention the conditions under which they can be considered to be justified.

We assume that the observed signal is a degraded version of the clean signal, where the degradation is classically modeled as the combination of linear channel distortion and additive noise [1]. The flow chart of the basic front-end signal processing is shown in Figure 12.2. Denoting by $y[t]$ the observed speech, $x[t]$ the clean speech, $n[t]$ the noise signal and $h[t]$ the impulse response of the linear channel-distortion filter, we obtain the following relationship in the time domain, where $*$ denotes convolution:

$$y[t] = (h * x)[t] + n[t]. \tag{12.4}$$

The frequency content of the observed signal is then generally analyzed using the short-term *discrete Fourier transform* (DFT): overlapping frames of the signal are windowed and the DFT is computed, leading to the complex short-term spectrum. Let us denote by $Y_{t,f}$ (respectively, $X_{t,f}$ and $N_{t,f}$) the spectrum of the observed speech (respectively, the clean speech and the noise) at time frame t and frequency bin f, and by H_f the DFT of h

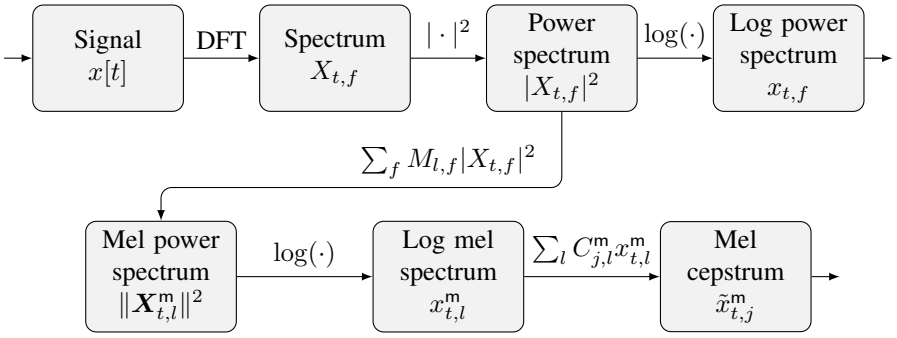


Figure 12.2 Basic front-end signal processing showing the notation used throughout the chapter for different feature domains.

(assumed shorter than the window length). Under the so-called narrowband approximation, the relationship between the complex short-term spectra can be written as:

$$Y_{t,f} \approx H_f X_{t,f} + N_{t,f}. \tag{12.5}$$

Note that this approximation can only be justified for a short channel-distortion filter and a smooth window function (i.e., whose Fourier transform is concentrated at low frequencies) [48]. This approximation is extremely common in frequency-domain source separation [42, 87].

We are now ready to transform (12.5) to the power spectrum domain:

$$|Y_{t,f}|^2 = |H_f|^2|X_{t,f}|^2 + |N_{t,f}|^2 + 2|H_f||X_{t,f}||N_{t,f}|\cos(\phi_{t,f}), \tag{12.6}$$

where $\phi_{t,f}$ is the phase difference between $H_f X_{t,f}$ and $N_{t,f}$. The third term is often assumed to be zero, leading to the following approximate interaction:

$$|Y_{t,f}|^2 \approx |H_f|^2|X_{t,f}|^2 + |N_{t,f}|^2. \tag{12.7}$$

This approximation is commonly justified by noticing that the expected value of the cross-term $|H_f||X_{t,f}||N_{t,f}|\cos(\phi_{t,f})$ is zero if x and n are assumed statistically independent. However, the expected value being equal to zero does not tell us much about the particular value taken at a given time-frequency bin. A slightly stronger argument to justify the above approximation is that of the sparsity of audio signals: if the speech and noise signals are sparse in the time-frequency domain, their cross-term is likely to be very small most of the time. Nonetheless, this term is not equal to zero in general, and we will see that the influence of the cross-term is actually very complex.

In order to reduce the influence of pitch (and thus reduce the within-class variance relative to the between-class variance when recognizing phonemes or sub-phonemes), the power spectrum is converted to the so-called mel power spectrum. The mel power spectrum is obtained by filtering the power spectrum using a small number L (typically 20 to 24 at a sampling rate of 8 kHz, 40 at 16 kHz) of overlapping triangular filters with both center frequencies and bandwidths equally spaced on the mel scale, believed to well approximate

the human perception of frequency. Denoting by $M_{l,f}$ the response of filter l in frequency f, the mel power spectrum of the observed signal is defined as

$$\|\boldsymbol{Y}_{t,l}^{\mathsf{m}}\|^2 = \sum_f M_{l,f}|Y_{t,f}|^2, \tag{12.8}$$

with similar definitions for that of the clean speech, $\|\boldsymbol{X}^{\mathsf{m}}\|^2$, and of the noise, $\|\boldsymbol{N}^{\mathsf{m}}\|^2$. As the number L of filters is typically much smaller than the number F of frequency bins, considering the mel power spectrum implies reducing the dimensionality of the features. Moreover, apart from reducing the influence of pitch, it also implicitly changes the weight given to the data as a function of frequency, in particular down-weighting the contribution of high frequencies. In terms of noise robustness, the mel domain has a beneficial effect for voiced speech in broadband noise: it gives preferential weight to the peaks of the spectrum, which are likely to correspond to the harmonics of speech, where the signal-to-noise ratio (SNR) is greatest. This is easy to see in the log mel domain, since $\log \sum_f M_{l,f}|Y_{t,f}|^2 \approx \max_f \log(M_{l,f}|Y_{t,f}|^2)$. Finally, we shall see that, as a side effect, it also leads to greater accuracy in the log-sum approximation, which is introduced in Section 12.4.3.

We can now obtain an analog of (12.6) on the mel spectra:

$$\|\boldsymbol{Y}_{t,l}^{\mathsf{m}}\|^2 = \|\boldsymbol{H}_{t,l}^{\mathsf{m}}\|^2\|\boldsymbol{X}_{t,l}^{\mathsf{m}}\|^2 + \|\boldsymbol{N}_{t,l}^{\mathsf{m}}\|^2 + 2\sqrt{\|\boldsymbol{H}_{t,l}^{\mathsf{m}}\|^2\|\boldsymbol{X}_{t,l}^{\mathsf{m}}\|^2\|\boldsymbol{N}_{t,l}^{\mathsf{m}}\|^2}\,\alpha_{t,l}^{\mathsf{m}}, \tag{12.9}$$

where the two newly introduced quantities

$$\|\boldsymbol{H}_{t,l}^{\mathsf{m}}\|^2 = \frac{\sum_f M_{l,f}|H_f|^2|X_{t,f}|^2}{\|\boldsymbol{X}_{t,l}^{\mathsf{m}}\|^2} \tag{12.10}$$

$$\alpha_{t,l}^{\mathsf{m}} = \frac{\sum_f M_{l,f}|H_f||X_{t,f}||N_{t,f}|\cos(\phi_{t,f})}{\sqrt{\|\boldsymbol{H}_{t,l}^{\mathsf{m}}\|^2\|\boldsymbol{X}_{t,l}^{\mathsf{m}}\|^2\|\boldsymbol{N}_{t,l}^{\mathsf{m}}\|^2}} \tag{12.11}$$

incorporate complex interactions between the various terms. These terms are typically not handled using the above formulae but through approximate models, as shown later in this chapter.

In order to deal with the very wide dynamic range of speech, and motivated by considerations on the roughly logarithmic perception of loudness by humans, the power spectrum and mel power spectrum are often converted to the log domain. We define the log power spectrum of the observed signal as

$$y_{t,f} = \log(|Y_{t,f}|^2) \tag{12.12}$$

with analogous definitions for the log power spectra of the clean speech $x_{t,f}$, the noise $n_{t,f}$, and the channel distortion h_f. This leads to the following interaction function in the log power domain:

$$y_{t,f} = \log\left(\mathrm{e}^{h_f + x_{t,f}} + \mathrm{e}^{n_{t,f}} + 2\mathrm{e}^{\frac{h_f + x_{t,f} + n_{t,f}}{2}}\cos(\phi_{t,f})\right). \tag{12.13}$$

Similarly to the log power spectrum, we define the log mel power spectrum of the noisy speech as

$$y_{t,l}^{\mathsf{m}} = \log(\|\boldsymbol{Y}_{t,l}^{\mathsf{m}}\|^2) \tag{12.14}$$

and analogously for the log power spectra of the clean speech, $x^{\mathsf{m}}_{t,l}$, the noise, $n^{\mathsf{m}}_{t,l}$, and the channel distortion, $h^{\mathsf{m}}_{t,l}$. The interaction function in the log mel power domain becomes

$$y^{\mathsf{m}}_{t,l} = \log\left(\mathrm{e}^{h^{\mathsf{m}}_{t,l}+x^{\mathsf{m}}_{t,l}} + \mathrm{e}^{n^{\mathsf{m}}_{t,l}} + 2\mathrm{e}^{\frac{h^{\mathsf{m}}_{t,l}+x^{\mathsf{m}}_{t,l}+n^{\mathsf{m}}_{t,l}}{2}}\alpha^{\mathsf{m}}_{t,l}\right). \tag{12.15}$$

To decorrelate the features, and focus on the envelope characteristics of the log mel power spectrum, which are likely to be related to the characteristics of the vocal tract, most speech recognition systems further compute the so-called mel cepstrum, which consists of the low-frequency components of the discrete cosine transform (DCT) of the log mel power spectrum. Introducing the DCT matrix $\mathbf{C}^{\mathsf{m}}$ of size $K \times L$, where K is the number of mel cepstral coefficients (typically around 13), the mel cepstrum of the noisy signal is defined as

$$\tilde{\boldsymbol{y}}^{\mathsf{m}}_t = \mathbf{C}^{\mathsf{m}}\boldsymbol{y}^{\mathsf{m}}_t, \tag{12.16}$$

with similar definitions of the mel cepstra of the clean speech, $\tilde{\boldsymbol{x}}^{\mathsf{m}}_t$, the noise, $\tilde{\boldsymbol{n}}^{\mathsf{m}}_t$, and the channel distortion, $\tilde{\boldsymbol{h}}^{\mathsf{m}}_t$. As the matrix $\mathbf{C}^{\mathsf{m}}$ is typically not invertible, the interaction function in the mel cepstrum domain is generally approximated by

$$\tilde{\boldsymbol{y}}^{\mathsf{m}}_t = \mathbf{C}^{\mathsf{m}}\log\left(\mathrm{e}^{\mathbf{D}^{\mathsf{m}}(\tilde{\boldsymbol{h}}^{\mathsf{m}}_t+\tilde{\boldsymbol{x}}^{\mathsf{m}}_t)} + \mathrm{e}^{\mathbf{D}^{\mathsf{m}}\tilde{\boldsymbol{n}}^{\mathsf{m}}_t} + 2\mathrm{e}^{\mathbf{D}^{\mathsf{m}}\frac{\tilde{\boldsymbol{h}}^{\mathsf{m}}_t+\tilde{\boldsymbol{x}}^{\mathsf{m}}_t+\tilde{\boldsymbol{n}}^{\mathsf{m}}_t}{2}} \circ \boldsymbol{\alpha}^{\mathsf{m}}_t\right), \tag{12.17}$$

where $\mathbf{D}^{\mathsf{m}}$ is a pseudoinverse of $\mathbf{C}^{\mathsf{m}}$, such as the Moore–Penrose pseudoinverse, and $\circ$ denotes the element-wise product.

Notice that the interaction becomes more and more complicated and nonlinear as we move closer to the features that are used in modern speech recognizers. When we consider using probabilistic models of the speech and noise in the feature domain, the more complicated the interaction function is, the less tractable inference becomes. To make matters worse, state-of-the-art systems do not stop at the mel cepstrum, but introduce further transformations that encompass multiple frames. These include linear transformations of several frames of features, such as the so-called *delta* and *delta-delta* features and *linear discriminant analysis* (LDA) as well as nonlinear transformations such as *feature-based minimum phone error* (fMPE). Except for the simplest cases, model-based noise compensation with such features has not yet been addressed. We shall thus limit our presentation mainly to *static* (i.e., single frame) features.

12.4 Interaction Models

For each of the feature domains introduced above, we have shown that a domain-specific interaction function describes how noisy features relate to those of the clean speech and the additive noise. In feature domains such as the complex spectrum, which contain complete information about the underlying signals, the interaction function is deterministic. However, in feature domains that omit some information, the unknown information leads to uncertainty about the interaction, which in the model-based approach is described using a probabilistic interaction function. In general, the model-based approach thus requires a distribution $p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{n}_t)$ over the observed noisy features $\boldsymbol{y}_t$ given the speech features $\boldsymbol{x}_t$ and the noise features $\boldsymbol{n}_t$. The definition of this function varies depending on the feature domain. In the

feature domains most used for modeling speech, approximations are generally required to make inference tractable. In this section, we review interaction models for log spectrum features, as well as some of their extensions to the mel spectrum domain and the mel cepstrum. From here on we omit time subscripts to simplify notation, bearing in mind that we are modeling the interaction in a particular time frame t.

12.4.1 Exact Interaction Model

We consider for now the modeling of speech and noise energy in the log power spectrum domain. In (12.13), the unknown phase and channel are a source of uncertainty in the relationship between the power spectra of the speech and noise. In the log power spectrum, the effect of the acoustic channel is well approximated as an additive constant, for stationary reverberation with an impulse response of length less than a frame. We can thus model the channel implicitly as part of the speech feature, to simplify our discussion, with little loss in generality. See [13] for a review in which it is explicitly included in the interaction model. The difference in phase ϕ_f between the speech and noise signals is a remaining source of uncertainty

$$p(y_f|x_f, n_f, \phi_f) = \delta\left(y_f - \log\left(\mathrm{e}^{x_f} + \mathrm{e}^{n_f} + 2\mathrm{e}^{\frac{x_f+n_f}{2}}\cos(\phi_f)\right)\right). \tag{12.18}$$

We need to compute $\int_{-\pi}^{\pi} p(y_f|x_f, n_f, \phi_f)p(\phi_f)\,\mathrm{d}\phi_f$. We define $\alpha_f = \cos(\phi_f)$ and derive $p_{\alpha_f}(\alpha_f)$ from $p_{\phi_f}(\phi_f)$, noting that $\cos(\phi_f) = \cos(-\phi_f)$, so that for $\phi_f \in (-\pi, \pi)$, we have two solutions to $|\phi_f| = \cos^{-1}(\alpha_f)$:

$$p_{\alpha_f}(\alpha_f) = \frac{p_{\phi_f}(\phi_f) + p_{\phi_f}(-\phi_f)}{\left|\frac{\partial\cos(\phi_f)}{\partial\phi_f}\right|} = \frac{p_{\phi_f}(\cos^{-1}(\alpha_f)) + p_{\phi_f}(-\cos^{-1}(\alpha_f))}{\sqrt{1-\alpha_f^2}}. \tag{12.19}$$

Given a distribution over α_f, the log spectrum interaction model can be written generally as

$$p(y_f|x_f, n_f) = p_{\alpha_f}(\alpha_f)\left|\frac{\partial y_f}{\partial\alpha_f}\right|^{-1} \tag{12.20}$$

$$= p_{\alpha_f}\left(\frac{1}{2}\left(\mathrm{e}^{y_f - \frac{x_f+n_f}{2}} - \mathrm{e}^{\frac{x_f-n_f}{2}} - \mathrm{e}^{\frac{n_f-x_f}{2}}\right)\right)\frac{1}{2}\mathrm{e}^{y_f - \frac{x_f+n_f}{2}}, \tag{12.21}$$

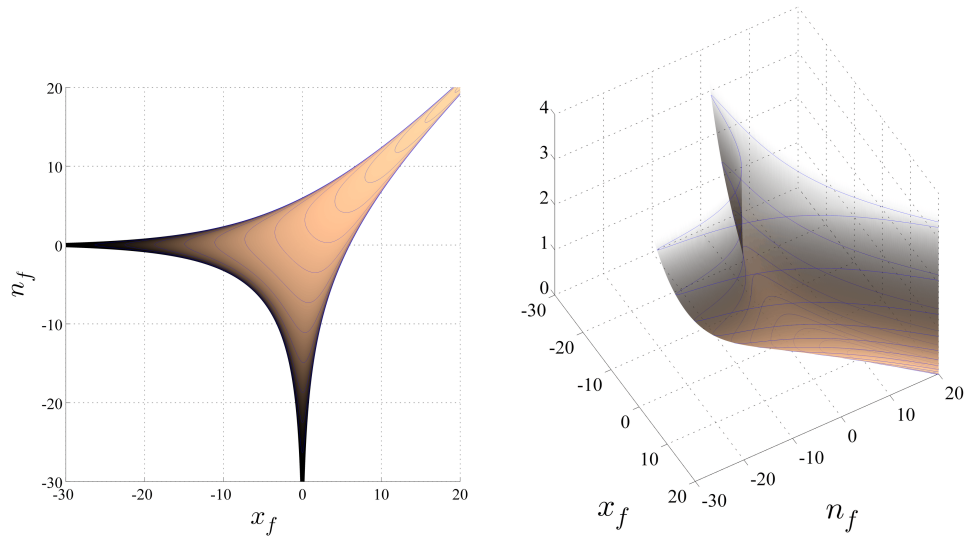
where α_f is obtained as a function of y_f, x_f and n_f from (12.13).

If we assume that the phase difference between speech and noise ϕ_f is uniformly distributed, $p_{\phi_f}(\phi_f) = \frac{1}{2\pi}$, then a change of variables leads to

$$p_{\alpha_f}(\alpha_f) = \frac{1}{\pi\sqrt{1-\alpha_f^2}}. \tag{12.22}$$

This is a shifted *beta distribution*: $p_{\alpha_f}(\alpha_f) = \frac{1}{2}\mathrm{Beta}\left(\frac{\alpha_f+1}{2}; a = \frac{1}{2}, b = \frac{1}{2}\right)$ where $\mathrm{Beta}(x; a, b) = x^{a-1}(1-x)^{b-1}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ [23]. Indicating uniform phase by $\mathcal{U}(\phi)$, we substitute (12.22) yielding:

$$p_{\mathcal{U}(\phi)}(y_f|x_f, n_f) = \frac{\frac{1}{2\pi}\mathrm{e}^{y_f - \frac{x_f+n_f}{2}}}{\sqrt{1 - \frac{1}{4}\left(\mathrm{e}^{y_f - \frac{x_f+n_f}{2}} - \mathrm{e}^{\frac{x_f-n_f}{2}} - \mathrm{e}^{\frac{n_f-x_f}{2}}\right)^2}} \tag{12.23}$$



(a) Contour plot of exact interaction model

(b) Surface plot of exact interaction model

Figure 12.3 (a) Contour plot of the exact density in the log spectrum $p_{\mathcal{U}(\phi)}(y_f = 0|x_f, n_f)$ and (b) surface plot of the same, showing how the function is unbounded at the edges. Contour line spacing is logarithmic and the function has been truncated to fit in the plot box.

as shown in [37], where it is called the *devil function* after its tortuous shape. Note that interesting alternate expressions for the same quantity can be obtained after some algebraic manipulations:

$$p_{\mathcal{U}(\phi)}(y_f|x_f, n_f)$$

$$= \frac{\frac{1}{\pi}\mathrm{e}^{y_f}}{\sqrt{\left(\mathrm{e}^{\frac{y_f}{2}} + \mathrm{e}^{\frac{x_f}{2}} + \mathrm{e}^{\frac{n_f}{2}}\right)\left(-\mathrm{e}^{\frac{y_f}{2}} + \mathrm{e}^{\frac{x_f}{2}} + \mathrm{e}^{\frac{n_f}{2}}\right)\left(\mathrm{e}^{\frac{y_f}{2}} - \mathrm{e}^{\frac{x_f}{2}} + \mathrm{e}^{\frac{n_f}{2}}\right)\left(\mathrm{e}^{\frac{y_f}{2}} + \mathrm{e}^{\frac{x_f}{2}} - \mathrm{e}^{\frac{n_f}{2}}\right)}} \tag{12.24}$$

$$= \frac{\frac{1}{\pi}\mathrm{e}^{y_f}}{\sqrt{\left(\mathrm{e}^{y_f} + \mathrm{e}^{x_f} + \mathrm{e}^{n_f}\right)^2 - 2\left(\mathrm{e}^{2y_f} + \mathrm{e}^{2x_f} + \mathrm{e}^{2n_f}\right)}}. \tag{12.25}$$

Later, we discuss application of similar derivations to the mel domain considered in [90]. In the amplitude domain, a similar distribution is known in the wireless communications literature as the *two-wave envelope* pdf [21].

Figures 12.3(a) and 12.3(b) show the exact interaction density function (12.23). The interaction density is highly nonlinear, and diverges to infinity along the edges of the feasible region. The edge toward the bottom left of Figure 12.3(a), where $x_f < 0$ and $n_f < 0$, results from cases where the phase difference is zero and the signal amplitudes add up to the observation. The two other edges, where $x_f > 0$ or $n_f > 0$, result from cases where the signals have opposing phase and cancel to generate the observed signal.

Unfortunately, with (12.23), the integral in (12.1) is generally intractable, leaving sampling as the only viable approach for inference (see for example [37]). Therefore, there have been a series of approaches based on approximate interaction functions, especially in the mel domain, to which we will turn after discussing more basic approximations in the log spectrum domain.

12.4.2 Max Model

Approximating the sum of two signals in a frequency band as the maximum of the two signals is an intuitive idea that roughly follows our knowledge of masking phenomena in human hearing[2], and can be justified mathematically. Expressing (12.13) in the form:

$$y_f = \max(x_f, n_f) + \log\left(1 + \mathrm{e}^{-|x_f - n_f|} + 2\mathrm{e}^{-\frac{|x_f - n_f|}{2}} \cos(\phi_f)\right), \tag{12.26}$$

we can see that when one signal dominates the other, the second term approaches zero, taking the effect of phase along with it. This motivates the *max approximation*:

$$y_f \approx \max(x_f, n_f), \tag{12.27}$$

which can be interpreted probabilistically using a Dirac delta:

$$p_{\max}(y_f|x_f, n_f) \stackrel{\text{def}}{=} \delta\big(y_f - \max(x_f, n_f)\big). \tag{12.28}$$

Note that more general models based on the max approximation could be defined by additionally modeling the uncertainty associated with the approximation. For example, the approximation error could be modeled as Gaussian, and, optionally, made dependent on SNR. Such modeling has been thoroughly investigated for the log-sum approximation, as described below, but, to the best of our knowledge, has not yet been investigated for the max approximation.

Remarkably, the max approximation is the mean of the exact interaction function (12.23) [66][3]:

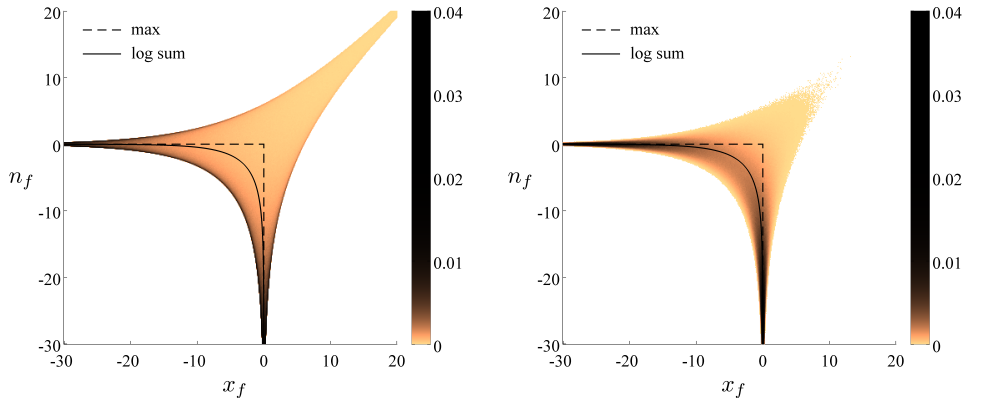
$$\mathrm{E}(y_f|x_f, n_f) = \int y_f \; p_{\mathcal{U}(\phi)}(y_f|x_f, n_f) \; \mathrm{d}y_f = \max(x_f, n_f). \tag{12.29}$$

The max approximation was first used for noise compensation in [62]. Shortly thereafter, in [91], it was used to compute joint state likelihoods of speech and noise and find their optimal state sequence using a factorial hidden Markov model.

Inference in the max model is generally intractable when $p(\boldsymbol{x}|s^{\mathsf{x}})$ or $p(\boldsymbol{n}|s^{\mathsf{n}})$ have dependencies across frequency, as do, for example, full-covariance Gaussians. However, for

[2] A high intensity signal at a given frequency affects the human hearing threshold for other signals at that frequency (signals roughly 6 dB below the dominant signal are not heard), and nearby frequencies, with diminishing effect, as a function of frequency difference. Consult [59] for details.

[3] While [66] reverts to an integration table to complete the proof of (12.29), it can be shown from (12.26) by noticing that $\forall \eta \in [0, 1)$, $\int \log(1 + \eta^2 + 2\eta\cos(\theta)) \; \mathrm{d}\theta = \int \log|1 + \eta \mathrm{e}^{\mathrm{i}\theta}|^2 \; \mathrm{d}\theta = 2\mathrm{Re}(\int \log(1 + \eta \mathrm{e}^{\mathrm{i}\theta}) \; \mathrm{d}\theta) = 0$, where θ is integrated over $[0, 2\pi)$. After a change of variable $z = \eta \mathrm{e}^{\mathrm{i}\theta}$, this can be obtained using Cauchy's integral formula $f(a) = \frac{1}{2\pi\mathrm{i}} \oint_\gamma \frac{f(z)}{z-a} \; \mathrm{d}z$ applied to the holomorphic function $f : z \mapsto \log(1 + z)$ defined in the open disk $\{z \in \mathbb{C} : |z| < 1\}$ of the complex plane, with $a = 0$ and on the circle γ of center 0 and radius η. The case $\eta = 1$ results from simple computations and amounts to showing $\int_0^\pi \log(\sin(\theta)) \; \mathrm{d}\theta = -\pi \log(2)$.



(a) Empirical interaction function for the log spectrum. (b) Empirical interaction function for an average of five power spectrum bins.

Figure 12.4 Histograms representing empirical measurements of the interaction function for (a) the log spectrum domain and (b) an average of five power spectrum bins typical of the mel spectrum domain.

conditionally independent models of the form $p(\boldsymbol{x}|s^{\mathsf{x}}) = \prod_f p(x_f|s^{\mathsf{x}})$, the state likelihoods and the posterior of $(\boldsymbol{x}, \boldsymbol{n})$ given the states can be readily computed, as shown below. Moreover, the max model is also highly amenable to approximate inference when explicitly evaluating all state combinations is computationally intractable, as described in Section 12.6.

12.4.3 Log-Sum Model

The *log-sum model*, used in [60, 27] based on the additivity assumption in the power domain [7], uses the log of the expected value in the power domain to define an interaction function:

$$y_f \approx \log \mathrm{E}(\mathrm{e}^{y_f}|x_f, n_f) = \log(\mathrm{e}^{x_f} + \mathrm{e}^{n_f}), \tag{12.30}$$

which can then be interpreted probabilistically using

$$p_{\mathrm{logsum}}(y_f|x_f, n_f) \stackrel{\mathrm{def}}{=} \mathcal{N}(y_f; \log(\mathrm{e}^{x_f} + \mathrm{e}^{n_f}), \psi_f), \tag{12.31}$$

where $\boldsymbol{\Psi}$ is a variance intended to compensate for the effects of phase. In the limit as $\psi_f \to 0$, $p_{\mathrm{logsum}}(y_f|x_f, n_f)$ becomes a Dirac delta function, leading to the model investigated in [20].

In the case of the log mel spectrum, which is closer to the features used by a recognizer, matters are made worse by the lack of a closed form expression for $p(y_l^{\mathsf{m}}|x_l^{\mathsf{m}}, n_l^{\mathsf{m}})$. This situation arises because the mel quantities are averages across frequency, but the signal interaction involves the whole frequency domain, as can be seen for example in (12.11). On the other hand, since the mel frequency domain averages together multiple bins, the effect of phase averages out. In this case, the log-sum approximation becomes more accurate, as shown in Figure 12.4(b).

However, the log-sum approximation does not account for the changing variance of y_f^{m} as a function of the SNR stemming from the complicated phase term in (12.9). Various approximations have been proposed to handle this [49, 15, 90, 86, 84].

12.4.4 Mel Interaction Model

Although directly integrating out phase in the mel spectrum interaction (12.15) is intractable, a frequently used approximation is to assume that the term α_l^{m} in (12.15) has a known distribution, $\tilde{p}(\alpha_l^{\mathsf{m}})$, that is independent of x_l^{m} and n_l^{m}. Using this approximation, we can directly use (12.21):

$$p_{\mathrm{mel}}(y_l^{\mathsf{m}}|x_l^{\mathsf{m}}, n_l^{\mathsf{m}}) \approx \tilde{p}_{\alpha_l^{\mathsf{m}}}\left(\frac{1}{2}\left(\mathrm{e}^{y_l^{\mathsf{m}} - \frac{x_l^{\mathsf{m}}+n_l^{\mathsf{m}}}{2}} - \mathrm{e}^{\frac{x_l^{\mathsf{m}}-n_l^{\mathsf{m}}}{2}} - \mathrm{e}^{\frac{n_l^{\mathsf{m}}-x_l^{\mathsf{m}}}{2}}\right)\right)\frac{1}{2}\mathrm{e}^{y_l^{\mathsf{m}} - \frac{x_l^{\mathsf{m}}+n_l^{\mathsf{m}}}{2}} \quad (12.32)$$

Unfortunately, it is still intractable to perform exact inference in this model. Hence, in [90], the integrals in (12.1) are computed by Monte Carlo, using a truncated Gaussian approximation to $\tilde{p}(\alpha_l^{\mathsf{m}})$. The shifted beta distribution mentioned earlier also has the feature that it can approximate a Gaussian for parameters a and b such that $ab \gg 1$, so perhaps it could be used as a unifying distribution, with empirically trained parameters, to handle the full range of cases. Approximate inference methods are discussed in Section 12.5.

12.5 Inference Methods

We have defined a number of interaction models, and now turn to inference methods for these interaction models. The main quantity of interest for speech recognition is the state likelihood $p(\boldsymbol{y}|s^{\mathsf{x}}, s^{\mathsf{n}})$ defined in (12.1). The posterior distribution of speech and noise $p(\boldsymbol{x}, \boldsymbol{n}|\boldsymbol{y}, s^{\mathsf{x}}, s^{\mathsf{n}})$ defined in (12.2) is also important but can often be computed using the same approximation methods as the likelihood. Figures 12.6 and 12.7 show how different the likelihoods can be for the various approximate inference methods described in this section.

12.5.1 Max Model Inference

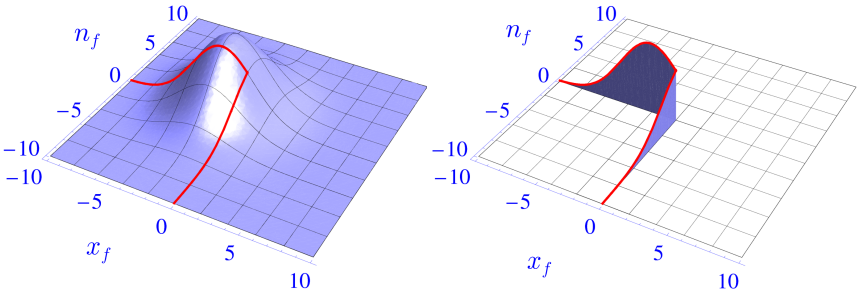
The likelihood of the speech and noise features, $\boldsymbol{x}$ and $\boldsymbol{n}$, under the max model is:

$$\begin{aligned} p_{\max}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n}) &= \prod_f p_{\max}(y_f|x_f, n_f) \\ &= \prod_f \delta(y_f - \max(x_f, n_f)). \end{aligned} \quad (12.33)$$

For models of the form $p(\boldsymbol{x}|s^{\mathsf{x}}) = \prod_f p(x_f|s^{\mathsf{x}})$ with conditionally independent features given the states (e.g., diagonal-covariance Gaussians), the state likelihoods and the posterior of $\boldsymbol{x}$ given the states can be readily computed. Define the probability density of the event $\mathsf{x}_f = y_f$ given state s^{x} as $p_{\mathsf{x}_f}(y_f|s^{\mathsf{x}})$, and the probability of the event that $\mathsf{x}_f \leq y_f$ as $\Phi_{\mathsf{x}_f}(y_f|s^{\mathsf{x}}) \stackrel{\text{def}}{=} p(\mathsf{x}_f \leq y_f) = \int_{-\infty}^{y_f} p(x_f|s^{\mathsf{x}})\mathrm{d}x_f$, which is the cumulative distribution function (cdf) of x_f given s^{x} evaluated at y_f.

For a given combination of states $s^{\mathsf{x}}, s^{\mathsf{n}}$, the cdf of y_f under the model factors for independent sources [62, 76]:

$$\begin{aligned} p(\mathsf{y}_f \leq y_f|s^{\mathsf{x}}, s^{\mathsf{n}}) &= p(\max(\mathsf{x}_f, \mathsf{n}_f) \leq y_f|s^{\mathsf{x}}, s^{\mathsf{n}}) \\ &= p(\mathsf{x}_f \leq y_f, \mathsf{n}_f \leq y_f|s^{\mathsf{x}}, s^{\mathsf{n}}) \\ &= \Phi_{\mathsf{x}_f}(y_f|s^{\mathsf{x}})\Phi_{\mathsf{n}_f}(y_f|s^{\mathsf{n}}). \end{aligned} \quad (12.34)$$



(a) Prior density (b) Posterior density

Figure 12.5 Inference under the max interaction model for clean speech x_f and noise n_f, for a single combination of states. In (a) the conditional prior $p(x_f, n_f|s^{\mathsf{x}}, s^{\mathsf{n}}) = p(x_f|s^{\mathsf{x}})p(n_f|s^{\mathsf{n}})$ is shown for a single feature dimension. The support of the likelihood function $p_{\max}(y_f|x_f, n_f) = \delta\left(y_f - \max(x_f, n_f)\right)$, for $y_f = 0$, is represented by a thick contour. The state likelihood $p(y_f = 0|s^{\mathsf{x}}, s^{\mathsf{n}})$ is the integral along this contour. The feature posterior $p(x_f, n_f|s^{\mathsf{x}}, s^{\mathsf{n}}, y_f = 0)$, which is proportional to the product of the prior and likelihood functions, is shown in (b).

The density of y_f is obtained by differentiating the cdf:

$$\begin{aligned} p(y_f|s^{\mathsf{x}}, s^{\mathsf{n}}) &= \frac{\mathrm{d}}{\mathrm{d}y_f}\left(\Phi_{\mathsf{x}_f}(y_f|s^{\mathsf{x}})\Phi_{\mathsf{n}_f}(y_f|s^{\mathsf{n}})\right) \\ &= p_{\mathsf{x}_f}(y_f|s^{\mathsf{x}})\Phi_{\mathsf{n}_f}(y_f|s^{\mathsf{n}}) + p_{\mathsf{n}_f}(y_f|s^{\mathsf{n}})\Phi_{\mathsf{x}_f}(y_f|s^{\mathsf{x}}). \end{aligned} \tag{12.35}$$

The density of $\boldsymbol{y}$ then is:

$$p(\boldsymbol{y}|s^{\mathsf{x}}, s^{\mathsf{n}}) = \prod_f \left(p_{\mathsf{x}_f}(y_f|s^{\mathsf{x}})\Phi_{\mathsf{n}_f}(y_f|s^{\mathsf{n}}) + p_{\mathsf{n}_f}(y_f|s^{\mathsf{n}})\Phi_{\mathsf{x}_f}(y_f|s^{\mathsf{x}})\right). \tag{12.36}$$

Inference under this model is illustrated in Figure 12.5, and compared to other methods in Figures 12.6 and 12.7.

In the case that the $p(\boldsymbol{x}|s^{\mathsf{x}})$ or $p(\boldsymbol{n}|s^{\mathsf{n}})$ have dependencies across frequency, such as with full-covariance Gaussians, inference in the max model is generally intractable. When the conditional joint cdf of $\boldsymbol{y}$ is differentiated with respect to each dimension of $\boldsymbol{y}$, we obtain an expression having 2^F terms:

$$p(\boldsymbol{y}|s^{\mathsf{x}}, s^{\mathsf{n}}) = \sum_{\mathcal{F}\in\mathcal{P}([1..F])} \frac{\partial\Phi_{\mathsf{x}}(\boldsymbol{y}|s^{\mathsf{x}})}{\partial\{y_f\}_{f\in\mathcal{F}}} \frac{\partial\Phi_{\mathsf{n}}(\boldsymbol{y}|s^{\mathsf{n}})}{\partial\{y_{f'}\}_{f'\in\bar{\mathcal{F}}}}, \tag{12.37}$$

where $\mathcal{F} \subset [1..F]$ is any subset of the feature dimensions, $\bar{\mathcal{F}}$ is its complement, and the power set $\mathcal{P}([1..F])$ is the set of all such subsets. A set $\mathcal{F}$ of feature indices corresponds to a hypothesis that $\mathsf{x}_f > \mathsf{n}_f$, $f \in \mathcal{F}$, or in other words that x dominates in the selected frequency bands. When computing these quantities we would typically start with the joint pdf for each

source, and integrate to obtain the term of interest:

$$\frac{\partial \Phi_{\mathsf{x}}(\boldsymbol{y}|s^{\mathsf{x}})}{\partial \{y_f\}_{f\in\mathcal{F}}} = \int_{\mathcal{R}_{\bar{\mathcal{F}}}} p_{\mathsf{x}}(\boldsymbol{y}_{\mathcal{F}}, \boldsymbol{y}_{\bar{\mathcal{F}}}|s^{\mathsf{x}}) dy_{\bar{\mathcal{F}}}, \tag{12.38}$$

where we denote a subset of the variables indexed by set $\mathcal{F}$ as $\boldsymbol{y}_{\mathcal{F}} = \{y_f\}_{f\in\mathcal{F}}$, and the region of integration is the negative half-space of $\boldsymbol{y}_{\bar{\mathcal{F}}}$ defined by $\mathcal{R}_{\bar{\mathcal{F}}} = \bigotimes_{f\in\bar{\mathcal{F}}}(-\infty, y_f]$. These integrals are intractable, in general, for conditionally dependent models. Such integrals are also used in the marginalization approach to missing data methods as discussed in Chapter 14, and are a source of difficulty in applying these methods in the cepstral domain.

The equations above can be directly generalized to the case of multiple independent sources, as shown in [76]. In the general case of conditionally dependent features, there are then K^F terms in the conditional pdf of $\boldsymbol{y}$, where K is the number of source signals. In the case of conditionally independent features, the model factorizes over frequency, and only univariate forms of the integrals above have to be computed. However, there remains an exponential number of combinations of the states of each source that need to be considered. Approximate techniques for addressing this computational issue are discussed below in the section on efficient inference methods.

12.5.2 Parallel Model Combination

In an approach known as *parallel model combination* (PMC), [28] makes use of the log-sum approximation, and assumes that the conditional probability $p_{\mathrm{pmc}}(y_f|s^{\mathsf{x}}, s^{\mathsf{n}})$ is a normal distribution in the log spectrum or log mel spectrum domain. Moment-matching is then used in the power domain to estimate the parameters of $p_{\mathrm{pmc}}(y_f|s^{\mathsf{x}}, s^{\mathsf{n}})$. To avoid clutter, we omit conditioning on the states and simply write $p_{\mathrm{pmc}}(y_f)$ in this section. For simplicity, we present the method using diagonal-covariance models. The method is straightforward to extend to the case where the models are full-covariance [28], or are defined in a transformed domain such as the mel cepstrum, although at considerable additional computational cost. PMC defines $p_{\mathrm{pmc}}(y_f) = \mathcal{N}(y_f; \hat{\mu}_{\mathsf{y}_f}, \hat{\sigma}_{\mathsf{y}_f})$, and chooses the mean $\hat{\mu}_{\mathsf{y}_f}$ and the variance $\hat{\sigma}_{\mathsf{y}_f}$ so that

$$\begin{aligned} \mathrm{E}_{\mathcal{N}(y_f;\hat{\mu}_{\mathsf{y}_f},\hat{\sigma}_{\mathsf{y}_f})}(|Y_f|^2) &= \mathrm{E}(|X_f|^2) + \mathrm{E}(|N_f|^2) \\ \mathrm{Var}_{\mathcal{N}(y_f;\hat{\mu}_{\mathsf{y}_f},\hat{\sigma}_{\mathsf{y}_f})}(|Y_f|^2) &= \mathrm{Var}(|X_f|^2) + \mathrm{Var}(|N_f|^2). \end{aligned} \tag{12.39}$$

As $x_f \sim \mathcal{N}(x_f; \mu_{\mathsf{x}_f}, \sigma_{\mathsf{x}_f})$, the following identities hold:

$$\begin{aligned} \mathrm{E}(|X_f|^2) &= \mathrm{E}(\mathrm{e}^{x_f}) = \mathrm{e}^{\mu_{\mathsf{x}_f} + \frac{1}{2}\sigma_{\mathsf{x}_f}} \\ \mathrm{Var}(|X_f|^2) &= \mathrm{Var}(\mathrm{e}^{x_f}) = (\mathrm{e}^{\sigma_{\mathsf{x}_f}} - 1)\mathrm{e}^{2\mu_{\mathsf{x}_f} + \sigma_{\mathsf{x}_f}}, \end{aligned} \tag{12.40}$$

and similarly for n_f. These identities can be inverted for y_f to yield

$$\begin{aligned} \hat{\mu}_{\mathsf{y}_f} &= \log \mathrm{E}_{\mathcal{N}(y_f;\hat{\mu}_{\mathsf{y}_f},\hat{\sigma}_{\mathsf{y}_f})}(|Y_f|^2) - \frac{1}{2}\hat{\sigma}_{\mathsf{y}_f} \\ \hat{\sigma}_{\mathsf{y}_f} &= \log\left(1 + \frac{\mathrm{Var}_{\mathcal{N}(y_f;\hat{\mu}_{\mathsf{y}_f},\hat{\sigma}_{\mathsf{y}_f})}(|Y_f|^2)}{(\mathrm{E}_{\mathcal{N}(y_f;\hat{\mu}_{\mathsf{y}_f},\hat{\sigma}_{\mathsf{y}_f})}(|Y_f|^2))^2}\right). \end{aligned} \tag{12.41}$$

Substituting (12.39) and then (12.40) into (12.41) yields

$$\begin{aligned}\hat{\mu}_{\mathsf{y}_f} &= \log \frac{\mathrm{e}^{\mu_{\mathsf{x}_f}+\frac{1}{2}\sigma_{\mathsf{x}_f}} + \mathrm{e}^{\mu_{\mathsf{n}_f}+\frac{1}{2}\sigma_{\mathsf{n}_f}}}{\mathrm{e}^{\frac{1}{2}\hat{\sigma}_{\mathsf{y}_f}}} \\ \hat{\sigma}_{\mathsf{y}_f} &= \log\left(1 + \frac{(\mathrm{e}^{\sigma_{\mathsf{x}_f}}-1)\mathrm{e}^{2\mu_{\mathsf{x}_f}+\sigma_{\mathsf{x}_f}} + (\mathrm{e}^{\sigma_{\mathsf{n}_f}}-1)\mathrm{e}^{2\mu_{\mathsf{n}_f}+\sigma_{\mathsf{n}_f}}}{(\mathrm{e}^{\mu_{\mathsf{x}_f}+\frac{1}{2}\sigma_{\mathsf{x}_f}} + \mathrm{e}^{\mu_{\mathsf{n}_f}+\frac{1}{2}\sigma_{\mathsf{n}_f}})^2}\right).\end{aligned} \tag{12.42}$$

In other words, PMC assumes that the distributions of the clean speech and the noise are log-normal, and approximates the sum of two log-normal distribution as another log-normal distribution whose parameters are estimated by moment-matching in the power domain. This method is known as the Fenton–Wilkinson method [24]. Returning to writing state-conditional models, and with the parameters of $p_{\mathrm{pmc}}(y_f|s^{\mathsf{x}}, s^{\mathsf{n}})$ in hand, the state likelihood can now be evaluated. Note that this method does not supply an estimate of the speech features given the noisy features.

PMC is the result of three approximations: the log-sum approximation, the assumption that $p(y_f|s^{\mathsf{x}}, s^{\mathsf{n}})$ is Gaussian in the log domain, and the Fenton–Wilkinson approximation, which uses moment-matching in the power domain instead of moment-matching in the log domain. The latter is problematic because the mean and variance in the power domain are not sufficient statistics of a log-normal distribution. Because of this, the mean and variance of the true conditional distribution $p(y_f|s^{\mathsf{x}}, s^{\mathsf{n}})$ in the log domain are generally different from those estimated by the Fenton–Wilkinson method, as can be seen in Figure 12.6, where the Fenton–Wilkinson approximation is compared to Monte-Carlo approximations of the true conditional distribution. A Monte-Carlo method known as *data-driven PMC* was developed in [28, 30] to address this problem. Data-driven PMC estimates the mean of $\boldsymbol{y}$ by sampling from the prior distributions of speech and noise, and computing the empirical mean of the noisy speech under the log-sum approximation. Other log-normal approximation methods for the sum of independent log-normal distributions have been proposed which instead directly estimate the sufficient statistics in the log domain [82, 69, 95]. Section 12.5.3 concerns another method that in some cases abandons the assumption that $p(y_f|s^{\mathsf{x}}, s^{\mathsf{n}})$ is log-normal altogether.

12.5.3 Vector Taylor Series Approaches

Unlike PMC, the vector Taylor series (VTS)-based approaches do not assume that the conditional probability distribution $p(y_f|s^{\mathsf{x}}, s^{\mathsf{n}})$ is Gaussian in the log spectrum or log mel spectrum domain. Instead they linearize the log-sum interaction function (12.31) about an expansion point that is optimized for each observed y_f. The resulting conditional probability distribution is non-Gaussian and performs better in general than the PMC approximation. As an added benefit, the method yields estimates of the clean speech and is amenable to feature-based and model-based methods. The early VTS work of [60] was further developed by completing the probabilistic framework and introducing iterations on the expansion point in an algorithm known as Algonquin [26, 50], which we describe here. Although the original algorithms included reverberation of the speech in the framework, we here relegate these channel components to the speech model for simplicity.

Here, we present the algorithm for general full-covariance models, and omit the dependency on states of each model for simplicity of notation. Note that for diagonal-covariance models, the features decouple and can be handled using the formula below

y_f

Figure 12.6 Comparison of the probability distribution $p(y_f|s^{\mathsf{x}}, s^{\mathsf{n}})$ under Gaussian priors for the speech and noise for different interaction models and inference methods. In all cases, the speech prior has mean 4 dB, and standard deviation 1 dB, and the noise prior has mean 0 dB and standard deviation 10 dB. The *MC uniform phase* is a Monte-Carlo approximation to the exact interaction model in the log spectral domain with uniform phase (the devil function), (12.21), using the max-model control variate method of [37]. *MC Gaussian phase factor* is the Monte-Carlo approximation to (12.53) with α variance 0.2, using a similar control variate approach. Both Monte-Carlo estimates are here computed with 10 000 samples per value of y_f. The *max model* and *PMC (Fenton–Wilkinson)* approaches are straightforward, whereas VTS approaches depend upon the expansion point and iteration. Here, we show VTS expanded at the prior mean.

independently for each feature. To handle the joint posterior, we concatenate $\boldsymbol{x}$ and $\boldsymbol{n}$ to form the joint vector $\boldsymbol{z} = [\boldsymbol{x}^\top \boldsymbol{n}^\top]^\top$ and use the function $g(\boldsymbol{z}) = \log(\mathrm{e}^{\boldsymbol{x}} + \mathrm{e}^{\boldsymbol{n}})$, where the logarithm and exponents operate element-wise on $\boldsymbol{x}$ and $\boldsymbol{n}$. Using a first-order Taylor series expansion at the point $\boldsymbol{z}_0$, the conditional distribution $p_{\mathrm{logsum}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n})$ introduced in (12.31) is approximated as

$$p_{\mathrm{logsum}}(\boldsymbol{y}|\boldsymbol{z}) \approx p_{\mathrm{linear}}(\boldsymbol{y}|\boldsymbol{z}; \boldsymbol{z}_0) = \mathcal{N}(\boldsymbol{y}; g(\boldsymbol{z}_0) + \mathbf{J}_g(\boldsymbol{z}_0)(\boldsymbol{z} - \boldsymbol{z}_0), \boldsymbol{\Psi}) \tag{12.43}$$

where $\boldsymbol{\Psi} = (\psi_f)_f$ and $\mathbf{J}_g(\boldsymbol{z}_0)$ is the Jacobian matrix of g, evaluated at $\boldsymbol{z}_0$:

$$\mathbf{J}_g(\boldsymbol{z}_0) = \left.\frac{\partial g}{\partial \boldsymbol{z}}\right|_{\boldsymbol{z}_0} = \left[\mathrm{diag}(\tfrac{\partial g}{\partial \boldsymbol{x}}) \quad \mathrm{diag}(\tfrac{\partial g}{\partial \boldsymbol{n}})\right]\Big|_{\boldsymbol{x}_0, \boldsymbol{n}_0} = \left[\mathrm{diag}(\tfrac{1}{1+\mathrm{e}^{\boldsymbol{n}_0 - \boldsymbol{x}_0}}) \quad \mathrm{diag}(\tfrac{1}{1+\mathrm{e}^{\boldsymbol{x}_0 - \boldsymbol{n}_0}})\right]. \tag{12.44}$$

We assume that $\boldsymbol{x}$ and $\boldsymbol{n}$ are independent and Gaussian distributed when conditioning on the corresponding speech and noise states (which we here omit):

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\mathsf{x}}, \boldsymbol{\Sigma}_{\mathsf{x}}), \qquad p(\boldsymbol{n}) = \mathcal{N}(\boldsymbol{n}; \boldsymbol{\mu}_{\mathsf{n}}, \boldsymbol{\Sigma}_{\mathsf{n}}). \tag{12.45}$$

y_f

Figure 12.7 Comparison of the same probability distributions as Figure 12.6, but with different Gaussian priors for the speech and noise. Here, the speech prior has mean 2 dB, and standard deviation 1 dB, and the noise prior has mean 0 dB and standard deviation 2 dB. In this case, the prior is close to the point where the max model is less accurate (0 dB SNR). PMC, on the other hand, appears to do better because the variances of speech and noise are closer to each other. VTS also is more accurate because the expansion point at the prior is closer to the posterior mode.

Hence, $\boldsymbol{z}$ is Gaussian distributed with mean and covariance

$$\boldsymbol{\mu}_{\mathsf{z}} = \begin{bmatrix} \boldsymbol{\mu}_{\mathsf{x}} \\ \boldsymbol{\mu}_{\mathsf{n}} \end{bmatrix}, \qquad \boldsymbol{\Sigma}_{\mathsf{z}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathsf{x}} & 0 \\ 0 & \boldsymbol{\Sigma}_{\mathsf{n}} \end{bmatrix}. \tag{12.46}$$

This leads to a simple linear Gaussian model with a Gaussian prior $p(\boldsymbol{z})$ and a Gaussian conditional distribution $p_{\mathrm{linear}}(\boldsymbol{y}|\boldsymbol{z};\boldsymbol{z}_0)$ whose mean is a linear function of $\boldsymbol{z}$ and whose covariance is independent of $\boldsymbol{z}$. It is an easy and classical result in Bayesian theory that both $p_{\mathrm{linear}}(\boldsymbol{y};\boldsymbol{z}_0)$ and the posterior $p_{\mathrm{linear}}(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{z}_0)$ are then Gaussian, and that their mean and covariance can be easily computed from those of the prior and the conditional distribution. In particular, we can obtain the mean and covariance of the posterior by completing the square with respect to $\boldsymbol{z}$ in the exponent of $p_{\mathrm{linear}}(\boldsymbol{y}|\boldsymbol{z};\boldsymbol{z}_0)p(\boldsymbol{z})$. The covariance $\boldsymbol{\Sigma}_{\mathsf{z}|\boldsymbol{y}}$ turns out to be independent of $\boldsymbol{y}$:

$$\boldsymbol{\Sigma}_{\mathsf{z}|\boldsymbol{y}} = \left[\boldsymbol{\Sigma}_{\mathsf{z}}^{-1} + \mathbf{J}_g(\boldsymbol{z}_0)^\top \boldsymbol{\Psi}^{-1} \mathbf{J}_g(\boldsymbol{z}_0)\right]^{-1}, \tag{12.47}$$

while the mean is given by

$$\boldsymbol{\mu}_{\mathsf{z}|\boldsymbol{y}} = \boldsymbol{\Sigma}_{\mathsf{z}|\boldsymbol{y}} \left[\boldsymbol{\Sigma}_{\mathsf{z}}^{-1}\boldsymbol{\mu}_{\mathsf{z}} + \mathbf{J}_g(\boldsymbol{z}_0)^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{y} - g(\boldsymbol{z}_0) + \mathbf{J}_g(\boldsymbol{z}_0)\boldsymbol{z}_0)\right] \stackrel{\text{def}}{=} \begin{bmatrix} \boldsymbol{\mu}_{\mathsf{x}|\boldsymbol{y}} \\ \boldsymbol{\mu}_{\mathsf{n}|\boldsymbol{y}} \end{bmatrix}. \tag{12.48}$$

Figure 12.8 Comparison of the probability distribution $p(y_f|s^{\mathrm{x}}, s^{\mathrm{n}})$ for VTS computed with different expansion points, using the same priors as Figure 12.6. Expansion points are a) the prior mean (*prior*) which is the most commonly used expansion point, b) the posterior mean, (*MMSE*) estimated using the *MC Gaussian phase factor* method shown in Figure 12.6, c) the point having maximum likelihood (*ML*) under the linearization, and d) the mode of the posterior distribution (*MAP*) on the log-sum approximation curve, computed by grid search. Note that the latter is discontinuous because it switches from one mode of the posterior to another which has greater posterior density, but less likelihood when integrated under the linearization.

By further integrating out $\boldsymbol{z}$ in $p_{\mathrm{linear}}(\boldsymbol{y}|\boldsymbol{z};\boldsymbol{z}_0)p(\boldsymbol{z})$, we obtain the mean and covariance of $p(\boldsymbol{y};\boldsymbol{z}_0)$:

$$\boldsymbol{\mu}_{\mathrm{y}} = g(\boldsymbol{z}_0) + \mathbf{J}_g(\boldsymbol{z}_0)(\boldsymbol{\mu}_{\mathrm{z}} - \boldsymbol{z}_0), \tag{12.49}$$

$$\boldsymbol{\Sigma}_{\mathrm{y}} = \boldsymbol{\Psi} + \mathbf{J}_g(\boldsymbol{z}_0)\boldsymbol{\Sigma}_{\mathbf{z}}\mathbf{J}_g(\boldsymbol{z}_0)^{\top}. \tag{12.50}$$

Note that although $p_{\mathrm{linear}}(\boldsymbol{y};\boldsymbol{z}_0)$ is Gaussian for a given expansion point, the value of $\boldsymbol{z}_0$ is the result of optimization and depends on $\boldsymbol{y}$ in a nonlinear way, so that the state likelihood is non-Gaussian as a function of $\boldsymbol{y}$.

We shall note as well that the posterior mean can be rewritten in a simpler and more intuitive way using the above covariance:

$$\boldsymbol{\mu}_{\mathbf{z}|\boldsymbol{y}} = \boldsymbol{\mu}_{\mathbf{z}} + \boldsymbol{\Sigma}_{\mathbf{z}}\mathbf{J}_g(\boldsymbol{z}_0)^{\top}\boldsymbol{\Sigma}_{\mathrm{y}}^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}_{\mathrm{y}}\right). \tag{12.51}$$

The posterior mean is thus obtained as the sum of the prior mean and a renormalized version of the bias between the observed noisy speech and the predicted value of the noisy speech at the prior mean given a linearization of the interaction function at $\boldsymbol{z}_0$.

The linearization point is important to the accuracy of the algorithm, as can be seen in Figure 12.8, and theoretically should be near the mode of the "true" posterior obtained using

Figure 12.9 Comparison of the probability distribution $p(y_f|s^{\mathsf{x}}, s^{\mathsf{n}})$ for iterative VTS at different iterations, using the same priors as Figure 12.6. The convergence properties of iterative VTS are shown by plotting each of the first 20 iterations, followed by each of the last 10 iterations for a total of 30. The fact that these last iterations still differ on the left-hand tail of the distribution indicates that the algorithm is oscillating between different solutions. Here, the iterations are started at the prior mean, but other expansion points lead to similar behavior. It is interesting to note that in this case, the minimum of the last several iterations of VTS makes a nice approximation of the probability distribution given by *MC Gaussian phase factor* shown in Figure 12.6.

$p_{\text{logsum}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n})$ as the conditional probability. Therefore, whereas the initial linearization point is at the prior mean, in each iteration the estimated posterior mean is used to obtain a new expansion point $\boldsymbol{z}_0 = \boldsymbol{\mu}_{\mathsf{z}|\boldsymbol{y}}$. Because the interaction function is shift invariant, in the sense that $\boldsymbol{y} + \boldsymbol{v} = g(\boldsymbol{x} + \boldsymbol{v}, \boldsymbol{n} + \boldsymbol{v})$ for any $\boldsymbol{v}$, the linearization at $\boldsymbol{z}_0 = [\boldsymbol{x}_0; \boldsymbol{n}_0]$ is a plane tangent to g along the line in $\boldsymbol{x}$, $\boldsymbol{n}$ defined by $\boldsymbol{x} - \boldsymbol{x}_0 = \boldsymbol{n} - \boldsymbol{n}_0$. Since $\boldsymbol{y}$ is observed, this is equivalent, as illustrated in Figure 12.10, to linearizing at a point on the curve determined by the observation $\boldsymbol{y} = g(\boldsymbol{x}'_0, \boldsymbol{n}'_0)$, defined by $\boldsymbol{x}'_0 = \boldsymbol{\mu}_{\mathsf{x}|\boldsymbol{y}} + \boldsymbol{v}$, $\boldsymbol{n}'_0 = \boldsymbol{\mu}_{\mathsf{n}|\boldsymbol{y}} + \boldsymbol{v}$, where $\boldsymbol{v} = \boldsymbol{y} - g(\boldsymbol{\mu}_{\mathsf{x}|\boldsymbol{y}}, \boldsymbol{\mu}_{\mathsf{n}|\boldsymbol{y}})$. This point is not necessarily at the posterior mode along the curve, so the expansion can be a source of trouble for the algorithm. Most notably, it does not guarantee the convergence of the likelihood estimate which is known to fail in many conditions [49], as illustrated in Figure 12.9. It may be better to pose the problem in terms of finding the mode of the posterior distribution directly. Optimization methods such as quasi-Newton methods involve differentiating the log posterior, and thus compute differentials of $g(\boldsymbol{x})$, but can step toward the optimum in a smoother and faster way [51].

Our discussion of VTS approaches above has assumed the use of source models based in the log (mel) power spectral domain, rather than cepstral domain, and neglected the explicit modeling of channel effects. Both circumstances can be readily handled in the VTS framework, assuming that an invertible DCT matrix $\mathbf{C}$ is used to transform to the cepstral

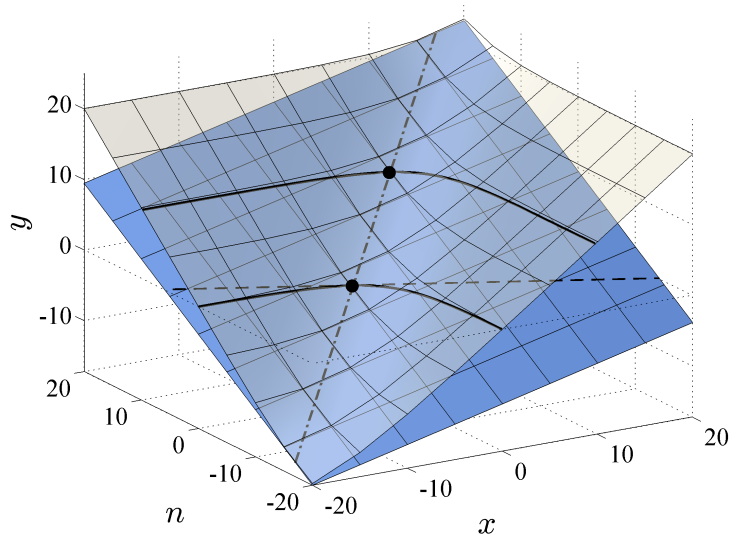
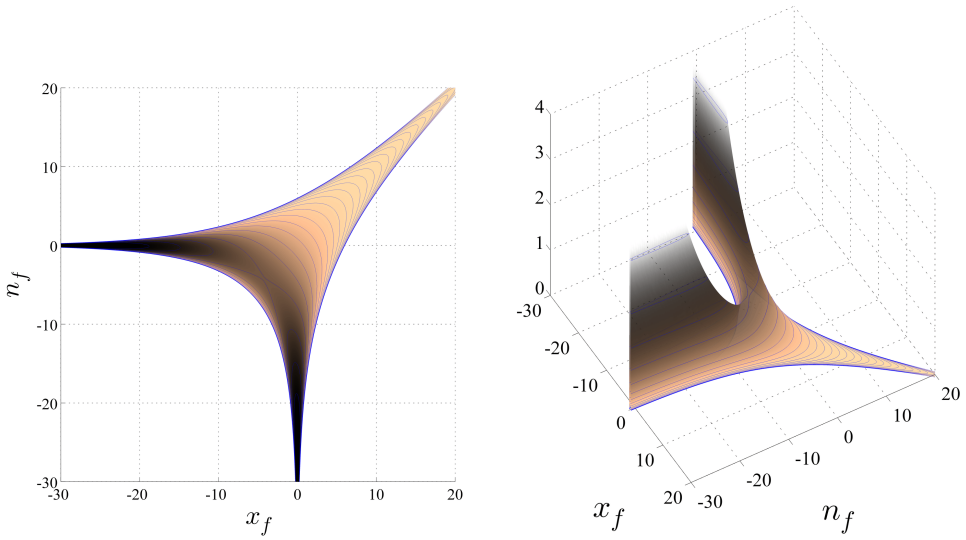


Figure 12.10 Illustration of the linearization procedure in VTS for a single frequency. The transparent surface on top represents the log-sum interaction $y = \log(\mathrm{e}^x + \mathrm{e}^n)$, while the plane below it is the linearization of g, which is tangent to that surface at $(x_0, n_0, g(x_0, n_0))$, for $(x_0, n_0) = (5, 8)$. This plane is also tangent along the dash-dotted line, because g has the property that $g(x + v, n + v) = y + v$ for any v. The two solid curves represent $y = \log(\mathrm{e}^x + \mathrm{e}^n)$ for $y = 0$ and $y = g(x_0, n_0)$. The dashed line is the tangent to $0 = \log(\mathrm{e}^x + \mathrm{e}^n)$ at (x_0, n_0) in the $y = 0$ plane.

domain. However, often the cepstra are generated by eliminating higher-order coefficients, in order to minimize the influence of pitch, and the Moore–Penrose pseudoinverse is commonly used. A more principled approach would be to supply a model of the upper cepstra so that the transformation is invertible.

In general, recognizers model features of multiple frames rather than a single one. This creates a model in which inference at the current frame is dependent upon previous frames. In [16], models of both *static* (i.e., single frame) and *dynamic* (i.e., differences across frame) features are used as priors for Algonquin. Although exact inference in such a model is generally intractable, [16] made the expedient approximation of using point estimates of the clean speech of previous frames to compute the priors of the current frame.

Unfortunately, as mentioned earlier, state-of-the-art speech recognizers use more complex and non-invertible transformations of multiple frames, such as LDA or fMPE transforms. Because of the nonlinearity and dimensionality reduction, further approximations would be necessary to perform model-based noise compensation with such models. In general, as previously mentioned, fleshing out the models to provide some distributions of the dimensions that are normally discarded is one avenue of attack.



(a) Contour plot of phase-factor model

(b) Surface plot of phase-factor model

Figure 12.11 (a) Contour plot of phase factor approximation and (b) surface plot of the same. Contour line spacing is logarithmic and the function has been truncated to fit in the plot box.

12.5.4 SNR-Dependent Approaches

SNR-dependent approaches [49, 20, 15], also known as "phase-sensitive" approaches, are similar to the basic VTS model except that a Gaussian model is used for the *phase factor* $\boldsymbol{\alpha}$ in (12.9), rather than for the entire phase term. Thus, neglecting the channel effects, the model is

$$\boldsymbol{y} = \log(\mathrm{e}^{\boldsymbol{x}} + \mathrm{e}^{\boldsymbol{n}} + 2\mathrm{e}^{\frac{\boldsymbol{x}+\boldsymbol{n}}{2}} \circ \boldsymbol{\alpha}). \tag{12.52}$$

The phase factor $\boldsymbol{\alpha} \in [-1, 1]^F$ is modeled as a zero mean Gaussian $p(\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{\alpha}; 0, \boldsymbol{\Sigma}_\alpha)$ truncated to the interval $[-1, 1]$ in each dimension. The variance $\boldsymbol{\Sigma}_\alpha$ is usually assumed to be diagonal. Using (12.21), we then have

$$p_{\mathrm{snrdep}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n}) = \mathcal{N}_\alpha\left(\frac{1}{2}(\mathrm{e}^{\boldsymbol{y}-\frac{\boldsymbol{x}+\boldsymbol{n}}{2}} - \mathrm{e}^{\frac{\boldsymbol{x}-\boldsymbol{n}}{2}} - \mathrm{e}^{\frac{\boldsymbol{n}-\boldsymbol{x}}{2}}); 0, \boldsymbol{\Sigma}_\alpha\right)\left|\mathrm{diag}\left(\frac{1}{2}\mathrm{e}^{\boldsymbol{y}-\frac{\boldsymbol{x}+\boldsymbol{n}}{2}}\right)\right|. \tag{12.53}$$

This distribution is illustrated in Figure 12.11. It is especially appropriate in the log mel domain and corresponds closely to the empirical distribution shown in Figure 12.4(b). Although the variance of α does not change as a function of SNR, the uncertainty of $\boldsymbol{y}$ given $\boldsymbol{x}$ and $\boldsymbol{n}$ becomes a function of SNR due to the nonlinearity of the interaction. In [49], in addition to modeling α as a Gaussian, the interaction was also approximated using

$$\begin{aligned}\boldsymbol{y} &= \log(\mathrm{e}^{\boldsymbol{x}} + \mathrm{e}^{\boldsymbol{n}}) + \log\left(1 + \frac{2}{\mathrm{e}^{\frac{\boldsymbol{x}-\boldsymbol{n}}{2}} + \mathrm{e}^{\frac{\boldsymbol{n}-\boldsymbol{x}}{2}}} \circ \boldsymbol{\alpha}\right) \\ &\approx \log(\mathrm{e}^{\boldsymbol{x}} + \mathrm{e}^{\boldsymbol{n}}) + \frac{2}{\mathrm{e}^{\frac{\boldsymbol{x}-\boldsymbol{n}}{2}} + \mathrm{e}^{\frac{\boldsymbol{n}-\boldsymbol{x}}{2}}} \circ \boldsymbol{\alpha}.\end{aligned} \tag{12.54}$$

Using this interaction function in (12.20) leads to a conditionally Gaussian likelihood function:

$$p_{\text{snrdepvar}}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{n}) = \mathcal{N}\left(\boldsymbol{y}; \log(\mathrm{e}^{\boldsymbol{x}} + \mathrm{e}^{\boldsymbol{n}}), \mathbf{A}^{\top}\boldsymbol{\Sigma}_{\alpha}\mathbf{A}\right). \tag{12.55}$$

The matrix $\mathbf{A} \stackrel{\text{def}}{=} \operatorname{diag}\big(2/(\mathrm{e}^{\frac{\boldsymbol{x}-\boldsymbol{n}}{2}} + \mathrm{e}^{\frac{\boldsymbol{n}-\boldsymbol{x}}{2}})\big)$, where division is defined element-wise, is a function of the SNR, $\boldsymbol{x} - \boldsymbol{n}$. In this case, the dependency of the uncertainty upon the SNR clearly appears in the variance, which reaches a maximum for an SNR of zero.

In [15], posteriors of clean speech and likelihoods of noisy speech were computed using (12.53), using an improved version of the VTS/Algonquin method, based on second-order expansion of the joint distribution, $p_{\text{snrdep}}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{n})p(\boldsymbol{x})p(\boldsymbol{n})$. The proposed algorithm was used to estimate the likelihoods of noisy speech, the posterior mean of the clean speech and to optimize the noise model given noisy speech.

12.6 Efficient Likelihood Evaluation in Factorial Models

Exact inference methods for robust ASR using factorial models require computing the joint state likelihood $p(\boldsymbol{y}|s^{\mathsf{x}}, s^{\mathsf{n}}) \equiv p(\boldsymbol{y}|s^{\mathsf{z}})$, introduced in (12.1), for all combinations of speech and noise states. Therefore, exact inference generally becomes computationally intractable when the number of state combinations is large. Efficient approximate inference naturally involves either reducing the amount of computation required to estimate $p(\boldsymbol{y}|s^{\mathsf{x}}, s^{\mathsf{n}})$, reducing the number of state combinations that are evaluated, or both.

12.6.1 Efficient Inference Using the Max Model

In Section 12.5.1, we showed that if the conditional prior distributions of speech and noise have no statistical dependencies between features, the joint likelihood of a given combination of speech and noise states under the max interaction model is given by:

$$p_{\max}(\boldsymbol{y}|s^{\mathsf{x}}, s^{\mathsf{n}}) = \prod_f \left(p_{\mathsf{x}_f}(y_f|s^{\mathsf{x}})\Phi_{\mathsf{n}_f}(y_f|s^{\mathsf{n}}) + p_{\mathsf{n}_f}(y_f|s^{\mathsf{n}})\Phi_{\mathsf{x}_f}(y_f|s^{\mathsf{x}})\right), \tag{12.56}$$

For K explicitly modeled acoustic sources, the result becomes

$$p(\boldsymbol{y}|\{s^k\}) = \prod_f \sum_k p_{\mathsf{x}_f^k}(y_f|s^k) \prod_{j\neq k} \Phi_{\mathsf{x}_f^j}(y_f|s^j), \tag{12.57}$$

where s^k denotes the acoustic state of the kth source x^k, and $\{s^k\}$ denotes $\{s^k\}_{k=1}^K = \{s^1, s^2, \ldots, s^K\}$, a particular configuration of the state variables of each source.

An advantageous property of this likelihood function is that it is composed of terms with factors that depend on the state of a single acoustic source. Therefore, the cost of computing these factors scales linearly with the number of acoustic sources that are explicitly modeled. However, exact inference using the max model requires that the product of sums in (12.36) be computed for every combination of states, which scales exponentially with the number of sources. This is true even for models in which the feature dimensions are conditionally independent given the states.

The joint likelihood (12.57) is often approximated to depend only on the acoustic model of a single source, for example, as done in [38], where $p(y_f|\{s^k\}) \approx p_{\mathsf{x}_f^i}(y_f|s^i), i =$

$\arg\max_k \mu_{s^k}$. This averts the cost of computing the cumulative distribution functions and the additions and multiplications in (12.57), but inference still scales exponentially with K, since the resulting likelihood function is, in general, different for every combination of states. In the case that all Gaussians in all acoustic models share the same variance at each dimension, the branch-and-bound algorithm in [80] can be applied to do an approximate search for the MAP state configuration, but this approach also has exponential worst-case complexity, and is not well suited for approximating the likelihoods of the states, because the upper bounds produced during the search are very loose.

Recently, a new variational framework for the max model was introduced [77, 75, 78]. The framework hinges on the observation that in each feature dimension, a latent hidden variable, corresponding to the identity of the source that explains the data in that dimension, is being integrated out in the sum in (12.57). Denoting the *mask variable* for feature f by d_f, and a particular choice of mask values for all of the features by $\{d_f\} \stackrel{\text{def}}{=} \{d_f\}_{f=1}^F$, we have

$$p(\boldsymbol{y}, \{d_f\}|\{s^k\}) = \prod_f p(y_f, \{d_f\}|\{s^k\}) \tag{12.58}$$

$$= \prod_f p_{\mathsf{x}_f^{d_f}}(y_f|s^{d_f}) \prod_{j \neq d_f} \Phi_{\mathsf{x}_f^j}(y_f|s^j), \tag{12.59}$$

where $\mathsf{x}_f^{d_f}$ and s^{d_f} denote the feature and state of the source that explains feature f. Note that $p(y_f, \{d_f\}|\{s^k\})$ is simply the product of the probability that source d_f explains the data, and all other source features have values less than the data. This *lifted max model* is derived more rigorously in [78], and explicitly models which source explains each feature dimension. The lifted max model has the special property that $p(\boldsymbol{y}, \{d_f\}|\{s^k\})$ factors over the acoustic sources, which immediately implies that if the mask values $\{d_f\}$ are known, inference of the acoustic sources decouples. Since $p(y_f, \{d_f\}|\{s^k\})$ factors over frequency, it also follows that if the state combination is known, then the inference of each d_f decouples from the others. In general, it is intractable to compute all possible acoustic masks (2^F), or all possible state combinations ($\prod_k |s^k|$), but these properties can be exploited using variational methods.

By Jensen's inequality, the log probability of the data under the lifted max model can be lower-bounded as follows:

$$\log p(\boldsymbol{y}) = \log \sum_{\{s^k\},\{d_f\}} p(\boldsymbol{y}, \{s^k\}, \{d_f\}) \tag{12.60}$$

$$\geq \sum_{\{s^k\},\{d_f\}} q(\{s^k\}, \{d_f\}) \log \frac{p(\boldsymbol{y}, \{s^k\}, \{d_f\})}{q(\{s^k\}, \{d_f\})} \stackrel{\text{def}}{=} \mathcal{L}, \tag{12.61}$$

for any probability distribution q on the states $\{s^k\}$ and masks $\{d_f\}$. The difference between (12.60) and (12.61) is the Kullback–Leibler (KL) divergence between the exact posterior under the model, $p(\{s^k\}, \{d_f\}|\boldsymbol{y})$, and $q(\{s^k\}, \{d_f\})$ [44]:

$$D(q_{\{\mathsf{s}_k\},\{\mathsf{d}_f\}} \,||\, p_{\{\mathsf{s}_k\},\{\mathsf{d}_f\}|\boldsymbol{y}}) = \log p(\boldsymbol{y}) - \mathcal{L}, \tag{12.62}$$

where we use the random variable notation for s_k and d_f to indicate that the divergence is only a function of their distribution and not their values. When $q(\{s^k\}, \{d_f\}) = p(\{s^k\}, \{d_f\}|\boldsymbol{y})$,

the bound is tight. By optimizing the variational parameters of $q(\{s^k\}, \{d_f\})$ to maximize the lower bound $\mathcal{L}$ in (12.61), we at the same time minimize (12.62). The resulting q distribution can be utilized as a surrogate for the true posterior $p(\{s^k\}, \{d_f\}|\boldsymbol{y})$, and used to make predictions. Because the joint distribution $p(\boldsymbol{y}, \{s^k\}, \{d_f\})$ factors, any form of $q(\{s^k\}, \{d_f\})$ that factors over both $\{s^k\}$ and $\{d_f\}$ makes optimizing the bound $\mathcal{L}$ in (12.61) *linear* in the number of sources K, the number of features F, and number of states $\sum_k |s^k|$. For example, if $q(\{s^k\}, \{d_f\}) = \prod_f q(d_f) \prod_k q(s^k)$, the bound $\mathcal{L}$ becomes:

$$\begin{aligned}
\mathcal{L} &= \sum_{\{s^k\},\{d_f\}} \prod_f q(d_f) \prod_k q(s^k) \log \frac{\prod_f p_{\mathsf{x}_f^{d_f}}(y_f|s^{d_f}) \prod_{j \neq d_f} \Phi_{\mathsf{x}_f^j}(y_f|s^j) \prod_k p(s^k)}{\prod_f q(d_f) \prod_k q(s^k)} \\
&= \sum_{f,k} \Big(q_{\mathsf{d}_f}(k) \sum_{s^k} q(s^k) \log p_{\mathsf{x}_f^k}(y_f|s^k) + (1 - q_{\mathsf{d}_f}(k)) \sum_{s^k} q(s^k) \log \Phi_{\mathsf{x}_f^k}(y_f|s^k) \Big) \\
&\quad + \sum_f H(q_{\mathsf{d}_f}) - \sum_k D(q_{\mathsf{s}_k} || \, p_{\mathsf{s}_k}),
\end{aligned} \tag{12.63}$$

as shown in [77], where $H(q_{\mathsf{d}_f}) = -\sum_{d_f} q(d_f) \log q(d_f)$ denotes the entropy of q_{d_f}. Clearly the bound can be computed without considering combinations of source states, or combinations of feature mask configurations, and so scales linearly with the number of sources, states per source, and feature dimension. Importantly, this implies that the chosen q distribution can also be iteratively inferred in time linear in these variables. As described in Section 12.7.2, these variational approximations and their extensions have been explored in the context of multi-talker speech recognition.

12.6.2 Efficient Vector-Taylor Series Approaches

To make inference in VTS-based systems more efficient, the following approximations are typically made:

- The noise is modeled by a single Gaussian to reduce the number of joint states to the number of states in the speech model, so that $|s^{\mathsf{z}}| = |s^{\mathsf{x}}|$, where $|s|$ denotes the number of discrete values that the state s can take.
- The data likelihood $p_{\text{linear}}(\boldsymbol{y}|s^{\mathsf{z}}; \boldsymbol{z}_0)$ is assumed to have a diagonal covariance matrix:

$$\begin{aligned}
&p_{\text{linear}}(\boldsymbol{y}|s^{\mathsf{z}}; \boldsymbol{z}_0) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_{\mathsf{y}|s^{\mathsf{z}}}, \boldsymbol{\Sigma}_{\mathsf{y}|s^{\mathsf{z}}}) \approx \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_{\mathsf{y}|s^{\mathsf{z}}}, \text{diag}(\boldsymbol{\Sigma}_{\mathsf{y}|s^{\mathsf{z}}})), \\
&\boldsymbol{\mu}_{\mathsf{y}|s^{\mathsf{z}}} = g(\boldsymbol{z}_0) + \mathbf{J}_g(\boldsymbol{z}_0)(\boldsymbol{\mu}_{\mathsf{z}|s^{\mathsf{z}}} - \boldsymbol{z}_0), \\
&\boldsymbol{\Sigma}_{\mathsf{y}|s^{\mathsf{z}}} = \boldsymbol{\Psi} + \mathbf{J}_g(\boldsymbol{z}_0) \boldsymbol{\Sigma}_{\mathsf{z}|s^{\mathsf{z}}} \mathbf{J}_g(\boldsymbol{z}_0)^\top.
\end{aligned}$$

  This reduces the cost of evaluating $p_{\text{linear}}(\boldsymbol{y}|s^{\mathsf{z}}; \boldsymbol{z}_0)$ by a factor of F, the dimension of $\boldsymbol{y}$.
- The approximation of the conditional likelihood $p_{\text{logsum}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n})$ as a Gaussian with mean linear in $\boldsymbol{x}$ and $\boldsymbol{n}$ is shared by sufficiently similar speech states

$$p_{\text{logsum}}(\boldsymbol{y}|\boldsymbol{z}) \approx p_{\text{linear}}(\boldsymbol{y}|\boldsymbol{z}, s^{\mathsf{z}}) \approx p_{\text{linear}}(\boldsymbol{y}|\boldsymbol{z}, s^{\mathsf{r}_{\mathsf{z}}}, s^{\mathsf{z}}) \tag{12.64}$$

where the state s^{r_z} is a "low-resolution" surrogate for the joint state s^{z}, and $|s^{\mathsf{r}_z}| \ll |s^{\mathsf{z}}|$. s^{r_z} is often referred to in the speech literature as a "regression class variable" [31]. Similarly, hierarchical acoustic models, which consist of multiple acoustic models trained at different model resolutions in terms of number of components can be used to compute surrogate likelihoods using VTS-methods while "searching" for probable state combinations.

The amount of computational savings brought by (12.64) depends on the specific approximations made, and several have been proposed [30]. Techniques such as joint uncertainty decoding (JUD) and VTS-JUD [57, 96], introduced in more detail in Chapter 17, have the advantage that only $|s^{\mathsf{r}_z}|$ sets of "compensation" parameters need to be computed, but the parameters of all $|s^{\mathsf{z}}|$ states of the acoustic model need to be transformed. Predictive CMLLR (PCMLLR) [31], conversely, implements model compensation via a feature transformation:

$$p_{\text{linear}}(\boldsymbol{y}|s^{\mathsf{z}}) \approx p_{\text{cmllr}}(\boldsymbol{y}|s^{\mathsf{z}}, s^{\mathsf{r}_z}) = |\mathbf{A}_{s^{\mathsf{r}_z}}| \mathcal{N}(\mathbf{A}_{s^{\mathsf{r}_z}}\boldsymbol{y} + \boldsymbol{b}_{s^{\mathsf{r}_z}}; \boldsymbol{\mu}_{\mathsf{x}|s^{\mathsf{x}}}, \boldsymbol{\Sigma}_{\mathsf{x}|s^{\mathsf{x}}}) \tag{12.65}$$

where $\mathbf{A}_{s^{\mathsf{r}_z}}$ and $\boldsymbol{b}_{s^{\mathsf{r}_z}}$ are estimated to minimize the KL divergence of $p_{\text{cmllr}}(\boldsymbol{y}|s^{\mathsf{z}}, s^{\mathsf{r}_z})$ from $p_{\text{linear}}(\boldsymbol{y}|s^{\mathsf{z}})$, and the Jacobian determinant $|\mathbf{A}_{s^{\mathsf{r}_z}}|$ ensures that the distribution in the right-hand side normalizes over $\boldsymbol{y}$. Note that the parameters of the speech model are not modified. Compared to model transformation methods that utilize diagonal-covariance approximations of $\boldsymbol{\Sigma}_{\mathsf{y}|s^{\mathsf{z}}}$, PCMLLR has the advantage that correlation changes in the feature vector can be modeled via $\mathbf{A}_{s^{\mathsf{r}_z}}$. Such modeling has been shown to improve ASR performance [30]. Another important advantage is that the PCMLLR model can be adapted in a straightforward manner like CMLLR [30].

The computational burden of computing likelihoods for all combinations of states in VTS models can also be alleviated using variational methods. A variational form of Algonquin was first discussed in [27], and is described in detail for the assumption of Gaussian posteriors for $\boldsymbol{x}$ and $\boldsymbol{n}$ in [49]. These algorithms iterate between computing linear approximation(s) of the log-sum function given the current estimate(s) of the speech and noise, and optimizing a variational lower bound on the resulting approximation to the probability of the data to update the speech and noise estimate(s) and acoustic likelihoods. The idea of conditioning the variational posterior on auxiliary state variables to control the number of masks that are inferred when doing inference in the max model [75, 78] could be similarly applied in the Algonquin (or VTS) framework to control the number of Gaussians used to approximate the posterior distribution of the features.

12.6.3 Band Quantization

Band quantization (BQ) is a technique that can be used to reduce the number of likelihoods that need to be computed per dimension for models with conditionally independent features. A *band-quantized Gaussian mixture model* (BQGMM) is a diagonal-covariance GMM that is constrained as follows. At each feature dimension f, an additional discrete random variable a_f^{x} is introduced, and the feature distribution is assumed to be Gaussian given a_f^{x}. The

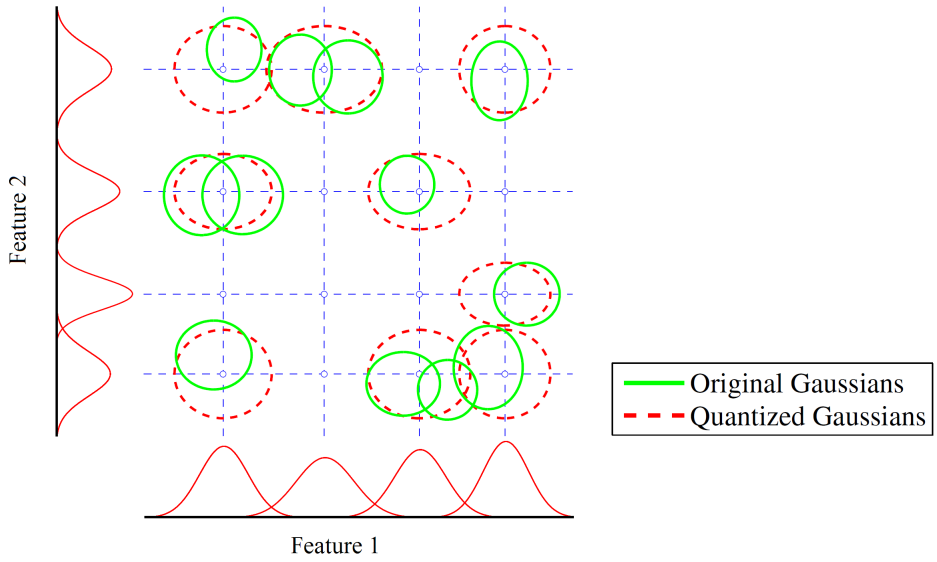


Figure 12.12 In band quantization, a large set of multidimensional Gaussians is represented using a small set of shared one-dimensional Gaussians optimized to best fit the original set of Gaussians. Here, we illustrate twelve two-dimensional Gaussians (solid ellipses). In each dimension, we quantize these to a pool of four shared one-dimensional Gaussians (density plots on axes). The means of these are drawn as a grid (dashed lines), on which the quantized two-dimensional Gaussians (dashed ellipses) can occur only at the intersections. Each quantized two-dimensional Gaussian is constructed from the corresponding pair of one-dimensional Gaussians, one for each feature dimension. In this example, we represent 24 means and variances (12 Gaussians $\times$ 2 dimensions), using 8 means and variances (4 Gaussians $\times$ 2 dimensions).

mapping from GMM states c^{x} to *atoms* a_f^{x} is usually constrained to be deterministic:

$$p(\boldsymbol{x}) = \sum_{c^{\mathsf{x}}} \pi_{c^{\mathsf{x}}} \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_{c^{\mathsf{x}}}, \boldsymbol{\sigma}^2_{c^{\mathsf{x}}}\right) \tag{12.66}$$

$$\approx \sum_{c^{\mathsf{x}}} \pi_{c^{\mathsf{x}}} \prod_f \mathcal{N}\left(x_f; \mu_{a_f^{\mathsf{x}}(c^{\mathsf{x}})}, \sigma^2_{a_f^{\mathsf{x}}(c^{\mathsf{x}})}\right). \tag{12.67}$$

By design $|a_f^{\mathsf{x}}| \ll |c^{\mathsf{x}}|$, so the number of Gaussians per dimension is vastly reduced. Figure 12.12 illustrates the idea. This concept was pioneered in early speech recognizers to reduce the computational load and promote generalization from small training sets [40, 5, 41, 6].

Despite the relatively small number of components $|a_f^{\mathsf{x}}|$ in each band, taken across bands, BQGMMs are capable of expressing $|a_f^{\mathsf{x}}|^F$ distinct patterns in an F-dimensional feature space. The computation and storage requirements of a BQGMM relative to its corresponding diagonal-covariance GMM are reduced by approximately a factor of $\frac{|c^{\mathsf{x}}|}{|a_f^{\mathsf{x}}|}$. For speech models,

this factor is generally on the order of 100 for negligible loss in ASR performance. The computational savings can be even more significant when using factorial models, which in general scale exponentially with the number of acoustic sources that are distinctly modeled. For example, in [38], BQGMMs are used to separate and recognize two simultaneously speaking talkers, and their use speeds up the cost of a full evaluation of the likelihoods by over three orders of magnitude. Importantly, band quantization can be applied to hierarchical acoustic models to reduce their memory footprint, and, depending on the search parameters used, deliver significant additional computational savings [4].

BQGMMs are generally estimated from an existing GMM, by clustering the Gaussians in each dimension using K-means clustering, with the KL-divergence between Gaussians as the distance metric [6]. More generally, BQGMMs can be identified by minimizing the KL-divergence between the BQGMM and an existing GMM. This objective cannot be analytically optimized. An analytic algorithm that uses variational techniques to approximate the KL divergence between GMMs is presented in [36], and was used to construct speech BQGMMs in [38].

12.7 Current Directions

We have reviewed some approaches to handling the problems of intractability in model-based approaches, both at the mathematical level, due to the nonlinearity of feature transforms, and at the computational level, due to the multiplicative number of state combinations for factorial models. We now discuss a few interesting current research directions in model-based robust ASR. The foregoing has focused on attempts to model signal interaction in feature domains that are known to work well for speech recognition. An alternative is to investigate speech recognition using feature domains in which signal interaction is easily modeled. Approaches to enhancement based on basis decomposition of power spectra attempt to model speech and noise directly in the power spectrum domain [81, 43, 74, 67].

Another direction is to investigate better modeling of speech and noise. The ability to model the noise dynamics is one of the more promising aspects of the model-based compensation framework. We discuss a model with simple linear dynamics on the noise levels that shows strong potential for use within a model-based noise compensation scheme. For more complex noise sources, such as an interfering speaker, noise compensation would be hopeless without complex models of the dynamics of both the target and interfering signals. However, some recent work on factorial HMMs shows that super-human speech recognition is possible and can be performed with far less computation than originally thought [78].

Speech recognition has a history that began with recognition of clean speech, and hence feature optimization has focused on extracting the filtering effects of the vocal tract and eliminating sources of variance that were thought irrelevant to recognition. The voiced parts of speech contain harmonics determined by the pitch, which carry the vocal tract information. However, in non-tonal languages the pitch is largely independent of the words being said. In noise, the situation changes: the harmonics are precisely the frequencies where the SNR is greatest, and so it may be profitable to model the dynamics of pitch along with the vocal tract information, in order to help extract the vocal tract information. Source-filter models also allow the interaction model to operate in the full spectrum, while allowing the recognition part of the model to operate in the filter domain. This type of model has been attempted for speech separation in [34, 54, 46], and for music separation in, for example, [33], and is also

used in HMM-based speech synthesis [97].

In the rest of this section, we discuss some of these ideas in more detail. In particular, we discuss dynamic noise models, speech separation with factorial HMMs, and non-negative subspace approaches to signal separation, and their potential use within a speech recognition system.

12.7.1 Dynamic Noise Models for Robust ASR

A fundamental problem in robust ASR (and classification in general) is handling mismatch between training and testing conditions in a highly efficient manner. Maximum Likelihood Linear Regression (MLLR) techniques such as fMLLR, Maximum a Posteriori Linear Regression (MAPLR), feature space MAPLR, etc. [28, 29, 10] are relatively simple, efficient and generally effective approaches to speaker and environmental compensation, and are used (in one form or another) by essentially all state-of-the-art ASR systems today.

However, as we have explained in detail in this chapter, additive noise has a highly nonlinear effect in the log frequency domain. Factorial models of speech and noise can exploit this relationship to learn efficient and representative models of the available training data. The benefits of explicitly modeling canonical variables such as noise are much more pronounced when mismatched data is encountered. Often very little adaptation data is available or very rapid adaptation is preferred. Naturally, an efficient and accurate parameterization of the data can be adapted much more rapidly and can be far more effective than brute-force methods.

The rapid adaptation of a noise model under a factorial representation of noisy speech is an idea with roots tracing back over four decades to early work on front-end denoising using spectral subtraction and Wiener filtering [7, 22]. Speech recognition systems are composed of loosely connected modules: a speech detector, a noise estimator that operates on blocks of data identified as speech-free, and a noise removal system, that produces a speech feature estimate given an estimate of the noise. Ongoing research aims to develop more accurate models of speech, noise, and their interaction, and jointly inferring their configuration under the resulting probability model of the data. More recent, significant work on rapid noise-adaptation includes investigations on dynamic forgetting factor algorithms for noise parameter adaptation [2], stochastic online noise parameter adaptation [14], and dynamic noise adaptation (DNA) [79, 71].

A distinguishing feature of DNA in this context is that noise is modeled as a random variable with simple dynamics. DNA maintains an approximation to the posterior distribution of the noise rather than a point estimate, which leads to better decisions about what frequency bands are explained by speech versus noise. A limitation of these rapid noise adaptation techniques is that they generally utilize very simple models of noise that are estimated in online fashion, and maintain no long-term statistics about previously seen data. The use of pre-trained models of noise to detect and reset the DNA noise tracker has been investigated to an extent [79], as has condition detection (CD): the automatic detection of when explicit noise modeling is not beneficial [72]. The latter approach allows for the use of DNA with multi-condition models for the speech model and back-end acoustic models, as is, without any system re-training, and improves the performance of state-of-the-art ASR systems significantly.

Factorial switching models with pre-trained (conditionally) linear dynamical models for speech and noise have also been investigated [18], and are described briefly in Chapter 9.

Future work on dynamic noise modeling should focus on efficiently leveraging stronger noise models that incorporate proven adaptation techniques, and incorporating/improving algorithms that have recently been investigated for multi-talker speech recognition (as described directly below), so that more structured acoustic interference, such as secondary speech and music, can be accurately compensated.

12.7.2 Multi-Talker Speech Recognition Using Graphical Models

A hallmark of human perception is our ability to solve the auditory cocktail party problem: even when restricted to a single channel, we can direct our attention to a chosen speaker in the presence of interfering speech, and, more often than not, understand what was said remarkably well. A truly exciting direction of current research in factorial modeling for robust ASR has been the use of graphical models to realize super-human speech recognition performance. These techniques have so far utilized HMMs to model each explicitly represented speaker, and combined them with one or more of the interaction models described in this chapter to realize multi-talker speech separation and recognition systems.

A fundamental challenge of multi-talker speech recognition is computational complexity. As discussed in Section 12.6, in general, exact inference involves computing the likelihood of all combinations of the states of the speakers. Exact inference also entails searching the joint (dynamic) state space of the decoders, which also scales exponentially with the number of speakers. In [35, 52, 38], the two-talker system used to outperform human listeners on the PASCAL monaural speech separation and recognition task [11], utilized band quantization (described in Section 12.6.3) to reduce the cost of acoustic labeling by an exponential factor, and joint-state pruning, which, for this well-constrained task, was very effective at controlling the complexity of the joint decoder. In [76] the idea of using loopy belief propagation to iteratively decode the speakers was introduced. This technique reduces the complexity of decoding from exponential to linear in the number of speakers, with negligible loss in recognition performance. Shortly thereafter in [77], the new variational framework for the max model described in Section 12.6.1 was introduced, and used to make inference linear in the number of speakers. Later in [75, 78], this framework was extended so that the complexity of inference could be precisely controlled. The resulting system was able to separate and recognize the speech of up to five speakers talking simultaneously and mixed in a single channel: a remarkable result, considering that the models necessary to describe the data involve trillions of state combinations for each frame.

These recent advances in multi-talker speech recognition are significant, but several important and exciting problems remain. First and foremost, it is important to emphasize that existing algorithms have so far only been tested in reasonably well-constrained scenarios, and artificially mixed data. The enhancement of these techniques to make them suitable for multi-talker recognition of real data streams with significant background noise, channel distortion, and less-constrained speaker vocabularies involves solving many interesting and challenging problems, some of which we discuss briefly below.

For example, to the best of our knowledge, algorithms that select which and how many speakers (or more generally acoustic sources) to explicitly model have yet to be investigated for more than two concurrently active sources. For the case of two sources, a simple method to detect clean conditions is described in [38]. This work, and existing work on speaker segmentation (e.g., [8]) could be used as a starting point for future investigations. Another

important direction of future work is to develop representative models of the acoustic background that extract canonical acoustic components that can be composed to explain new, previously unseen test data, and yet do not over-generalize. Current studies include matrix factorization approaches, as described further below, and factorial models based on graphical models with a distributed state representations, such as deep belief networks (DBNs) of restricted Boltzmann machines (RBMs) for ASR [58], and factorial hidden DBNs of RBMs for robust ASR [73].

In addition, relatively little work has been done on probabilistic models for speech separation and recognition that employ multiple channels in a coherent model[70, 3, 83, 12]. With the availability of two or more channels in speech enabled devices rapidly becoming the rule, rather than the exception, it seems inevitable that the best ASR systems will be those that have multi-channel processing capabilities integrated directly into the acoustic scorer and decoder.

12.7.3 Noise Robust ASR Using Non-Negative Basis Representations

We have so far shown how tremendous efforts need to be made in order to bring interaction modeling in the domain of the speech recognizer. Another approach to the problem is to try to perform recognition in a domain where the interaction can be conveniently modeled, such as the magnitude or power domain. A promising angle of attack in this direction is to use techniques based on non-negative matrix factorization (NMF) [55]. In the context of audio signal processing, NMF is generally applied to the magnitude or power spectrogram of the signal, with the hope that the non-negative low-rank decomposition thus obtained will extract relevant parts [88].

In NMF approaches, the model for each source in a given frame (a small window of speech, of approximately 40 ms) is defined by a set of weighted non-negative basis functions in the power spectrum (or similar feature space). Inference involves concatenating the basis sets for different sources into a single basis, and solving in parallel for the weights of all sources that best reconstruct the signal. There is also work to include phase explicitly as a parameter [45], which would allow for exact inference of the complete signal.

This type of approach has the advantage of speed because it avoids considering all combinations of basis functions across speakers. It has proven extremely successful, particularly for music signal transcription and source separation [92, 25], as described in more details in Chapter 5 of this book. The original framework has been extended in many directions. One has been to integrate better constraints, such as temporal continuity, into the models while retaining their computational advantages [92, 94]. Another has been to reformulate NMF in a probabilistic framework [9, 93], which enables posterior probabilities and likelihoods to be computed. This also enables NMF to be used as a component in a graphical model such as a speech HMM.

A recent trend of research has, like the speech separation approaches of the previous section, focused on modeling multiple non-stationary sources through factorial models [65, 89, 61, 64]. An exciting new direction takes this idea even further by using nonparametric Bayesian methods to define factorial models with an unbounded number of factors [39, 63]. The beauty of these methods is that, despite their apparent complexity, they are able to acquire models of all of the components of an acoustic scene. This makes them ideally suited to the task of modeling complex and unknown signals. Applying this kind of approach to the speech

recognition problem has, to the best of our knowledge, not yet been attempted, but we think that it is a very promising direction for future research.

References

[1] A. Acero. *Acoustical and Environmental Robustness in Automatic Speech Recognition*. Ph.D. thesis, Carnegie Mellon University, 1990.

[2] M. Afify and O. Siohan. Sequential noise estimation with optimal forgetting for robust speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 229–232, May 2001.

[3] H. Attias. New EM algorithms for source separation and deconvolution with a microphone array. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. 297–300, Apr. 2003.

[4] R. Bakis, D. Nahamoo, M. A. Picheny, and J. Sedivy. Hierarchical labeler in a speech recognition system, 2000. U.S. Patent 6023673.

[5] J. R. Bellegarda and D. Nahamoo. Tied mixture continuous parameter models for large vocabulary isolated speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 13–16, May 1989.

[6] E. Bocchieri. Vector quantization for the efficient computation of continuous density likelihoods. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 692–695, Apr. 1993.

[7] S. F. Boll. Suppression of acousic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(3):113–120, Apr. 1979.

[8] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair. Stream-based speaker segmentation using speaker factors and eigenvoices. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4133–4136, Apr. 2008.

[9] A. T. Cemgil. Bayesian inference in non-negative matrix factorisation models. Technical Report CUED/F-INFENG/TR.609, University of Cambridge, Jul. 2008.

[10] C. Chesta, O. Siohan, and C.-H. Lee. Maximum a posteriori linear regression for HMM adaptation. In *Proc. Interspeech ISCA European Conference on Speech Communication and Technology (Eurospeech)*, 1999.

[11] M. Cooke, J. R. Hershey, and S. J. Rennie. Monaural speech separation and recognition challenge. *Computer Speech and Language*, 24(1):1–15, Jan. 2010.

[12] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S.-J. Hahm, and N. A. Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation. In *Proc. International Workshop on Machine Listening in Multisource Environments (CHiME)*, Sep. 2011.

[13] L. Deng. Front-end, back-end, and hybrid techniques to noise-robust speech recognition. In D. Kolossa and R. Haeb-Umbach, editors, *Robust Speech Recognition of Uncertain or Missing Data Robust Speech Recognition of Uncertain or Missing Data*, chap. 4, pp. 67–99. Springer Verlag, 2011.

[14] L. Deng, J. Droppo, and A. Acero. Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 11(6):568–580, 2003.

[15] 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 Transactions on Speech and Audio Processing*, 12(2):133–143, 2004.

[16] L. Deng, J. Droppo, and A. Acero. Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features. *IEEE Transactions on Speech and Audio Processing*, 12(3):218–233, 2004.

[17] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer. Speaker adaptation using constrained estimation of Gaussian mixtures. *IEEE Transactions on Speech and Audio Processing*, 3(5):357–366, 1995.

[18] J. Droppo and A. Acero. Noise robust speech recognition with a switching linear dynamic model. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 953–956, May 2004.

[19] J. Droppo, A. Acero, and L. Deng. Uncertainty decoding with SPLICE for noise robust speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 57–60, May 2002.

[20] J. Droppo, L. Deng, and A. Acero. A comparison of three non-linear observation models for noisy speech features. In *Proc. Interspeech ISCA European Conference on Speech Communication and Technology (Eurospeech)*, pp. 681–684, 2003.

[21] G. D. Durgin, T. S. Rappaport, and D. A. D. Wolf. New analytical models and probability density functions for fading in wireless communications. *IEEE Transactions on Communications*, 50(6):1005–1015, 2002.

[22] ETSI. Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms. *ETSI TR ES 202 050 VERSION 1.1.3*, 2003.

[23] M. Evans, N. Hastings, and B. Peacock. *Statistical distributions*. Wiley-Interscience, 3rd edition, 2000.

[24] L. F. Fenton. The sum of lognormal probability distributions in scatter transmission systems. *IRE Transactions on Communication Systems*, CS-8:57–67, 1960.

[25] C. Févotte. Itakura-Saito nonnegative factorizations of the power spectrogram for music signal decomposition. In W. Wang, editor, *Machine Audition: Principles, Algorithms and Systems*, chap. 11. IGI Global Press, Aug. 2010.

[26] B. J. 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. Interspeech ISCA European Conference on Speech Communication and Technology (Eurospeech)*, pp. 901–904, Sep. 2001.

[27] B. J. Frey, T. T. Kristjansson, L. Deng, and A. Acero. ALGONQUIN - learning dynamic noise models from noisy speech for robust speech recognition. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14 (NIPS 2001)*, pp. 1165–1171, Cambridge, Massachusetts, 2002. MIT Press.

[28] M. J. F. Gales. *Model-based techniques for noise robust speech recognition*. Ph.D. thesis, University of Cambridge, Sep. 1995.

[29] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98, Jan. 1998.

[30] M. J. F. Gales. Model-based approaches to handling uncertainty. In D. Kolossa and R. Haeb-Umbach, editors, *Robust Speech Recognition of Uncertain or Missing Data Robust Speech Recognition of Uncertain or Missing Data*, chap. 5, pp. 101–125. Springer Verlag, 2011.

[31] M. J. F. Gales and R. C. van Dalen. Predictive linear transforms for noise robust speech recognition. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 59–64, 2007.

[32] J.-L. Gauvain and C.-H. Lee. Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, Apr. 1994.

[33] T. Heittola, A. Klapuri, and T. Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, Nov. 2009.

[34] J. R. Hershey and M. Casey. Audio-visual sound separation via hidden Markov models. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14 (NIPS 2001)*, pp. 1173–1180, Cambridge, Massachusetts, 2001. MIT Press.

[35] J. R. Hershey, T. T. Kristjansson, S. J. Rennie, and P. A. Olsen. Single channel speech separation using factorial dynamics. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19 (NIPS 2006)*, pp. 593–600. MIT Press, Cambridge, Massachusetts, 2007.

[36] J. R. Hershey and P. A. Olsen. Approximating the Kullback Leibler divergence between Gaussian mixture models. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2007.

[37] J. R. Hershey, P. A. Olsen, and S. J. Rennie. Signal interaction and the devil function. In *Proc. Interspeech ISCA International Conference on Spoken Language Processing (ICSLP)*, Sep. 2010.

[38] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson. Super-human multi-talker speech recognition: A graphical modeling approach. *Computer Speech and Language*, 24(1):45–66, Jan. 2010.

[39] M. D. Hoffman, D. M. Blei, and P. R. Cook. Bayesian nonparametric matrix factorization for recorded music. In *Proc. International Conference on Machine Learning (ICML)*, 2010.

[40] X. Huang and M. Jack. Semi-continuous hidden Markov models for speech signals. *Computer Speech & Language*, 3(3):239–251, 1989.

[41] M.-Y. Hwang. *Sub-phonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition*. Ph.D. thesis, Carnegie Mellon University, dec 1993.

[42] S. Ikeda and N. Murata. A method of ICA in time-frequency domain. In *Proc. International Conference on Independent Component Analysis and Blind Source Separation (ICA)*, pp. 365–370, 1999.

[43] G.-J. Jang and T.-W. Lee. A maximum likelihood approach to single-channel source separation. *Journal of Machine Learning Research*, 4:1365–1392, 2003.

[44] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, Nov. 1999.

[45] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama. Complex NMF: A new sparse representation for acoustic signals. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 3437–3440, Apr. 2009.

[46] H. Kameoka, N. Ono, and S. Sagayama. Speech spectrum modeling for joint estimation of spectral envelope and fundamental frequency. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1507–1516, Aug. 2010.

[47] D. K. Kim and M. J. F. Gales. Noisy constrained maximum likelihood linear regression for noise robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2):315–325, Feb. 2011.

[48] M. Kowalski, E. Vincent, and R. Gribonval. Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1818–1829, Sep. 2010.

[49] T. T. Kristjansson. *Speech Recognition in Adverse Environments*. Ph.D. thesis, University of Waterloo, 2002.

[50] T. T. Kristjansson, H. Attias, and J. R. Hershey. Single microphone source separation using high resolution signal reconstruction. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 817–820, May 2004.

[51] T. T. Kristjansson and R. Gopinath. Cepstrum domain Laplace denoising, 2005. *Available electronically at* http://www.research.ibm.com/people/r/rameshg/kristjansson-icassp2005.pdf.

[52] T. T. Kristjansson, J. R. Hershey, P. A. Olsen, S. J. Rennie, and R. Gopinath. Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system. In *Proc. Interspeech ISCA International Conference on Spoken Language Processing (ICSLP)*, 2006.

[53] H. Lane and B. Tranel. The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research*, 14(4):677, 1971.

[54] J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama. Single channel speech and background segregation through harmonic-temporal clustering. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 279–282, Oct. 2007.

[55] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, Oct. 1999.

[56] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of HMMs. *Computer Speech and Language*, 9:171–186, 1995.

[57] H. Liao and M. J. F. Gales. Joint uncertainty decoding for noise robust speech recognition. In *Proc. Interspeech ISCA International Conference on Spoken Language Processing (ICSLP)*, 2005.

[58] A. Mohamed, G. Dahl, and G. Hinton. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011 (to appear).

[59] B. C. J. Moore. *An Introduction to the Psychology of Hearing*. Emerald Group Publishing Ltd, 5th edition, 2003.

[60] P. J. Moreno, B. Raj, and R. M. Stern. A vector Taylor series approach for environment-independent speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 733–736, May 1996.

[61] G. J. Mysore, P. Smaragdis, and B. Raj. Non-negative hidden Markov modeling of audio with application to source separation. In *Proc. International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 140–148, 2010.

[62] A. Nádas, D. Nahamoo, and M. A. Picheny. Speech recognition using noise-adaptive prototypes. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(10):1495–1503, Oct. 1989.

[63] M. Nakano, J. Le Roux, H. Kameoka, T. Nakamura, N. Ono, and S. Sagayama. Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden Markov model. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2011.

[64] M. Nakano, J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama. Infinite-state spectrum model for music signal analysis. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1972–1975, May 2011.

[65] A. Ozerov, C. Févotte, and M. Charbit. Factorial scaled hidden Markov model for polyphonic audio representation and source separation. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2009.

[66] M. H. Radfar, A. H. Banihashemi, R. M. Dansereau, and A. Sayadiyan. Nonlinear minimum mean square error estimator for mixture-maximisation approximation. *Electronics Letters*, 42(12):724–725, Jun. 2006.

[67] B. Raj and P. Smaragdis. Latent variable decomposition of spectrograms for single channel speaker separation. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 17–20, Oct. 2005.

[68] B. Raj and R. M. Stern. Missing-feature approaches in speech recognition. *IEEE Signal Processing Magazine*, 22(5):101–116, Sep. 2005.

[69] C. K. Raut, T. Nishimoto, and S. Sagayama. Model composition by Lagrange polynomial approximation for robust speech recognition in noisy environment. In *Proc. Interspeech ISCA International Conference on Spoken Language Processing (ICSLP)*, pp. 2809–2812, 2004.

[70] S. J. Rennie, P. Aarabi, T. T. Kristjansson, B. Frey, and K. Achan. Robust variational speech separation using fewer microphones than speakers. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 88–91, Apr. 2003.

[71] S. J. Rennie, P. Dognin, and P. Fousek. Robust speech recognition using dynamic noise adaptation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2011.

[72] S. J. Rennie, P. Fousek, and P. Dognin. Matched-condition robust dynamic noise adaptation. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2011.

[73] S. J. Rennie, P. Fousek, and P. Dognin. Factorial hidden restricted Boltzmann machines for robust ASR. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012 (submitted).

[74] S. J. Rennie, J. R. Hershey, and P. A. Olsen. Efficient model-based speech separation and denoising using non-negative subspace analysis. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1833–1836, Apr. 2008.

[75] S. J. Rennie, J. R. Hershey, and P. A. Olsen. Hierarchical variational loopy belief propagation for multi-talker speech recognition. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 176–181, Dec. 2009.

[76] S. J. Rennie, J. R. Hershey, and P. A. Olsen. Single-channel speech separation and recognition using loopy belief propagation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 3845–3848, Apr. 2009.

[77] S. J. Rennie, J. R. Hershey, and P. A. Olsen. Variational loopy belief propagation for multi-talker speech recognition. In *Proc. Interspeech ISCA European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1331–1334, Sep. 2009.

[78] S. J. Rennie, J. R. Hershey, and P. A. Olsen. Single-channel multitalker speech recognition. *IEEE Signal Processing Magazine*, 27(6):66–80, 2010.

[79] S. J. Rennie, T. T. Kristjansson, P. A. Olsen, and R. Gopinath. Dynamic noise adaptation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 1197–1200, 2006.

[80] S. T. Roweis. Factorial models and refiltering for speech separation and denoising. In *Proc. Interspeech ISCA European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1009–1012, Sep. 2003.

[81] M. N. Schmidt and R. K. Olsson. Single-channel speech separation using sparse non-negative matrix factorization. In *Proc. Interspeech ISCA International Conference on Spoken Language Processing (ICSLP)*, pp. 2614–2617, Sep. 2006.

[82] S. Schwartz and Y. Yeh. On the distribution function and moments of power sums with lognormal components. *Bell System Technical Journal*, 61:1441–1462, 1982.

[83] M. L. Seltzer. *Microphone Array Processing for Robust Speech Recognition*. Ph.D. thesis, Carnegie Mellon University, 2003.

[84] M. L. Seltzer, A. Acero, and K. Kalgaonkar. Acoustic model adaptation via linear spline interpolation for robust speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4550–4553, 2010.

[85] M. L. Seltzer, B. Raj, and R. M. Stern. A Bayesian framework for spectrographic mask estimation for missing feature speech recognition. *Speech Communication*, 43(4):370–393, 2004.

[86] Y. Shinohara and M. Akamine. Bayesian feature enhancement using a mixture of unscented transformation for uncertainty decoding of noisy speech. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4569–4572, 2009.

[87] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22(1):21–34, 1998.

[88] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 177–180, 2003.

[89] P. Smaragdis and B. Raj. The Markov selection model for concurrent speech recognition. In *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 214–219, 2010.

[90] R. C. van Dalen and M. J. F. Gales. Asymptotically exact noise-corrupted speech likelihoods. In *Proc. Interspeech ISCA International Conference on Spoken Language Processing (ICSLP)*, 2010.

[91] A. P. Varga and R. K. Moore. Hidden Markov model decomposition of speech and noise. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 845–848, Apr. 1990.

[92] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, Mar. 2007.

[93] T. Virtanen, A. T. Cemgil, and S. Godsill. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1825–1828, Apr. 2008.

[94] K. W. Wilson, B. Raj, and P. Smaragdis. Regularized non-negative matrix factorization with temporal dependencies for speech denoising. In *Proc. Interspeech ISCA International Conference on Spoken Language Processing (ICSLP)*, pp. 411–414, Sep. 2008.

[95] J. Wu, N. B. Mehta, and J. Zhang. A flexible lognormal sum approximation method. In *Proc. IEEE Global Telecommunications Conference (Globecom)*, pp. 3413–3417, Dec. 2005.

[96] M. Xu, M. J. F. Gales, and K. K. Chin. Improving joint uncertainty decoding performance by predictive methods for noise robust speech recognition. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2009.

[97] H. Zen, K. Tokuda, and A. Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, 2009.

Index