Maas *et al. EURASIP Journal on Advances in Signal Processing* (2015) 2015:103
DOI 10.1186/s13634-015-0287-x

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

REVIEW Open Access

A Bayesian view on acoustic model-based techniques for robust speech recognition

Roland Maas[1*], Christian Huemmer[1], Armin Sehr[2] and Walter Kellermann[1]

Abstract

This article provides a unifying Bayesian view on various approaches for acoustic model adaptation, missing feature, and uncertainty decoding that are well-known in the literature of robust automatic speech recognition. The representatives of these classes can often be deduced from a Bayesian network that extends the conventional hidden Markov models used in speech recognition. These extensions, in turn, can in many cases be motivated from an underlying observation model that relates clean and distorted feature vectors. By identifying and converting the observation models into a Bayesian network representation, we formulate the corresponding compensation rules. We thus summarize the various approaches as approximations or modifications of the same Bayesian decoding rule leading to a unified view on known derivations as well as to new formulations for certain approaches.

Keywords: Robust automatic speech recognition, Bayesian network, Model adaptation, Missing feature, Uncertainty decoding

1 Introduction

Robust automatic speech recognition (ASR) still represents a challenging research topic. The main obstacle, namely the mismatch of test and training data, can be tackled by enhancing the observed speech signals or features in order to meet the training conditions or by compensating for the distorted test conditions in the acoustic model of the ASR system.

Methods that modify the acoustic model are in general termed (*acoustic*) *model-based* or *model compensation* approaches and comprise inter alia the following subcategories: so-called *model adaptation* techniques mostly update the parameters of the acoustic model, i.e., of the hidden Markov models (HMMs), prior to the decoding of a set of observed feature vectors. In contrast, *decoder-based* approaches re-adapt the HMM parameters for each observed feature vector. The most common decoder-based approaches are *missing feature* and *uncertainty decoding* that incorporate additional time-varying uncertainty information into the evaluation of the HMMs' probability density functions (pdfs).

Various model compensation techniques exhibit two (more or less) distinct steps: First, the compensation parameters need to be estimated and, second, the actual compensation rule is applied to the acoustic model. The compensation rules can often be motivated based on an observation model that relates the clean and distorted feature vectors, e.g., in the logarithmic melspectral (logmelspec) or the mel frequency cepstral coefficient (MFCC) domain.

In this article, we review and examine for several uncertainty decoding [1–5], missing feature [6–9], and model adaptation techniques [10–19] how their compensation rules can be formulated as an approximated or modified Bayesian decoding rule. In order to illustrate the formalism, we also present the corresponding Bayesian network representations. In addition to the above techniques, we give a Bayesian network description of the generic uncertainty decoding approach of [20], of the maximum a posteriori (MAP) adaptation technique [21], and of some alternative HMM topologies [22, 23]. While the Bayesian perspective [24, 25] and Bayesian networks have been employed in this context before [4, 9, 20], with this article, we present new formulations for certain of the considered algorithms in order to fill some gaps for a unified description. Throughout the following, feature

*Correspondence: maas@LNT.de
[1] Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg, Cauerstr. 7, 91058 Erlangen, Germany
Full list of author information is available at the end of the article

vectors are column vectors and denoted by bold-face letters $\mathbf{v}_n$ with time index $n \in \{1,\ldots,N\}$. Feature vector sequences are written as $\mathbf{v}_{1:N} = (\mathbf{v}_1,\ldots,\mathbf{v}_N)$. The operators "exp" and "log" applied to vectors are meant to be applied component-wise. The operator "$\odot$" denotes the component-wise vector multiplication (Hadamard product). Without distinguishing a random variable from its realization, a pdf over a random variable z_n is denoted by $p(z_n)$. For a normally distributed real-valued random vector $\mathbf{z}_n$ with mean vector $\boldsymbol{\mu}_{\mathbf{z}_n}$ and covariance matrix $\mathbf{C}_{\mathbf{z}_n}$, we write $\mathbf{z}_n \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}_n}, \mathbf{C}_{\mathbf{z}_n})$ or

$$p(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n; \boldsymbol{\mu}_{\mathbf{z}_n}, \mathbf{C}_{\mathbf{z}_n}). \tag{1}$$

To express that all random vectors of the set $\{\mathbf{z}_1,\ldots,\mathbf{z}_N\}$ share the same statistics, we write $p(\mathbf{z}_n) = \text{const.}$ or in the Gaussian case,

$$p(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n; \boldsymbol{\mu}_{\mathbf{z}}, \mathbf{C}_{\mathbf{z}}) \tag{2}$$

with time-invariant mean vector $\boldsymbol{\mu}_{\mathbf{z}}$ and covariance matrix $\mathbf{C}_{\mathbf{z}}$. Finally, for a Gaussian random vector $\mathbf{z}_n$ conditioned on another random vector $\mathbf{w}_n$, we write

$$p(\mathbf{z}_n|\mathbf{w}_n) = \mathcal{N}(\mathbf{z}_n; \boldsymbol{\mu}_{\mathbf{z}|\mathbf{w}_n}, \mathbf{C}_{\mathbf{z}|\mathbf{w}_n}), \tag{3}$$

if the statistics of $\mathbf{z}_n$ depend only on time through $\mathbf{w}_n$, i.e., if $\boldsymbol{\mu}_{\mathbf{z}|\mathbf{w}_n} = \boldsymbol{\mu}_{\mathbf{z}|\mathbf{w}_m}$ and $\mathbf{C}_{\mathbf{z}|\mathbf{w}_n} = \mathbf{C}_{\mathbf{z}|\mathbf{w}_m}$ for $\mathbf{w}_n = \mathbf{w}_m$ and $n, m \in \{1,\ldots,N\}$.

The remainder of the article is organized as follows: After summarizing the employed Bayesian view in Section 2 and its difference to other overview articles in Section 3, this perspective is applied to uncertainty decoding, missing feature techniques, and other model-based approaches in Sections 4, 5, and 6, respectively. In Section 7, we point out the relation of the presented techniques to deep learning-based architectures. Finally, conclusions are drawn in Section 8.

2 The Bayesian view

We start by reviewing the Bayesian perspective on acoustic model-based techniques that we use in Sections 4, 5, 6, and 7 to review different algorithms.

Given a sequence of observed feature vectors $\mathbf{y}_{1:N}$, the acoustic score $p(\mathbf{y}_{1:N}|\mathbf{W})$ of a sequence $\mathbf{w}$ of conventional HMMs, as depicted in Fig. 1a, is given by [24]

$$p(\mathbf{y}_{1:N}|\mathbf{W}) = \sum_{q_{1:N}} p(\mathbf{y}_{1:N}, q_{1:N}) \tag{4}$$

$$= \sum_{q_{1:N}} \left\{ \prod_{n=1}^{N} p(\mathbf{y}_n|q_n)\, p(q_n|q_{n-1}) \right\}, \tag{5}$$

where $p(q_1|q_0) = p(q_1)$. The summation goes over all possible state sequences $q_{1:N}$ through $\mathbf{W}$ superseding the explicit dependency on $\mathbf{w}$ at the right-hand side of (4) and (5). Note that the pdf $p(\mathbf{y}_n|q_n)$ can be scaled by $p(\mathbf{y}_n)$ without influencing the discrimination capability of the acoustic score w.r.t. changing word sequences $\mathbf{w}$. We thus define $\mathring{p}(\mathbf{y}_n|q_n) = p(\mathbf{y}_n|q_n)/p(\mathbf{y}_n)$ for later use.

The compensation rules of a wide range of model adaptation, missing feature, and uncertainty decoding approaches can be expressed by modifying the Bayesian network structure of a conventional HMM and applying the inference rules of Bayesian networks [26]—potentially followed by suitable approximations to ensure mathematical tractability. While some approaches postulate a certain Bayesian network structure, others indirectly define a modified Bayesian network by assuming an observed feature vector $\mathbf{y}_n$ to be a *distorted* version of an underlying *clean* feature vector $\mathbf{x}_n$, which is introduced as latent variable in the HMM as, e.g., in Fig. 1b. In the latter case, the relation of $\mathbf{y}_n$ and $\mathbf{x}_n$ can be expressed by an analytical observation model $g(\cdot)$ that incorporates certain compensation parameters $\mathbf{b}_n$:

$$\mathbf{y}_n = g(\mathbf{x}_n, \mathbf{b}_n). \tag{6}$$

Note that here it is not distinguished whether $\mathbf{y}_n$ is the output of a front-end enhancement process or a noisy or reverberant observation that is directly fed into the recognizer. By converting the observation model to a Bayesian network representation, the pdf $p(\mathbf{y}_n|q_n)$ in (5) can be derived exploiting the inference rules of Bayesian networks [26]. For the case of Fig. 1b, the observation likelihood in (5) would, e.g., become:

$$p(\mathbf{y}_n|q_n) = \int p(\mathbf{x}_n, \mathbf{y}_n|q_n) d\mathbf{x}_n$$

$$= \int p(\mathbf{x}_n|q_n)\, p(\mathbf{y}_n|\mathbf{x}_n) d\mathbf{x}_n, \tag{7}$$

where the actual functional form of $p(\mathbf{y}_n|\mathbf{x}_n)$ depends on the assumptions on $g(\cdot)$ and the statistics $p(\mathbf{b}_n)$ of $\mathbf{b}_n$.

The abstract perspective taken in this paper reveals a fundamental difference between model adaptation approaches on the one hand and missing feature and uncertainty decoding approaches on the other hand: Model adaptation techniques usually assume $\mathbf{b}_n$ to have constant statistics over time [4, 27], i.e.,

$$p(\mathbf{b}_n) = \text{const., for } n \in \{1,\ldots,N\}. \tag{8}$$

or to be a deterministic parameter vector of value $\mathbf{b}$, i.e.,

$$p(\mathbf{b}_n) = \delta(\mathbf{b}_n - \mathbf{b}), \tag{9}$$

where $\delta(\cdot)$ denotes the Dirac distribution. In contrast, missing feature and uncertainty decoding approaches typically assume $p(\mathbf{b}_n)$ to be a time-varying pdf [4, 27].

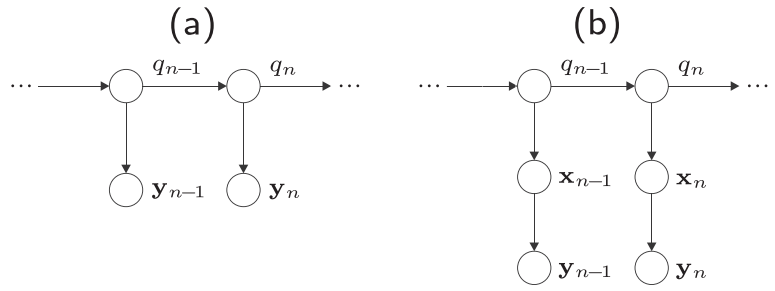

Fig. 1 Bayesian network representation of **a** a conventional HMM and **b** an HMM incorporating latent feature vectors. Subfigure **b** is based on [20]

As exemplified in Sections 4 to 6, this Bayesian view also allows for a convenient illustration of the underlying statistical dependencies of model-based approaches by means of Bayesian networks. If two approaches share the same Bayesian network, their underlying joint pdfs over all involved random variables share the same decomposition properties. However, some crucial aspects are not reflected by a Bayesian network: the particular functional form of the joint pdf, potential approximations to arrive at a tractable algorithm, as well as the estimation procedure for the compensation parameters. While some approaches estimate these parameters through an acoustic front-end, others derive them from clean or distorted data. For clarity, we entirely focus in this article on the compensation rules while ignoring the parameter estimation step. We also disregard approaches that apply a modified training method to conventional HMMs without exhibiting a distinct compensation step, as it is characteristic for, e.g., discriminative [28], multi-condition [29], or reverberant training [30].

3 Merit of the Bayesian view

In the past decades, numerous survey papers and books have been published summarizing the state-of-the-art in noise and reverberation-robust ASR [27, 31–36]. Recently, a comprehensive review of noise-robust ASR techniques was published in [25] providing a taxonomy-oriented framework by distinguishing whether, e.g., prior knowledge, uncertainty processing or an explicit distortion model is used or not. In contrast to [25] and previous survey articles, we pursue a threefold goal with this article:

- First of all, we aim at classifying all considered techniques along the same dimension by motivating and describing them with the same Bayesian formalism. Consequently, we do not conceptually distinguish whether a given method employs a time-varying pdf $p(\mathbf{b}_n)$, as in uncertainty decoding, or whether a distorted vector $\mathbf{y}_n$ is a preprocessed or a genuinely noisy or reverberant observation. Also, the distinction of implicit and explicit observation models dissolves in our formalism.
- As a second goal, we aim at closing some gaps by presenting new derivations and formulations for some of the considered techniques. For instance, the Bayesian networks in Figs. 2b, 2c, 4, 5b, 5c, 6b, 8, 9b representing the concepts in Subsections 4.3, 4.4, 4.6, 6.3/6.4, 6.5, 6.6, 6.8, and 6.9, respectively, constitute novel representations. Moreover, the links to the Bayesian framework via the mathematical reformulations in (28), (29), (37), (38), (45), (55), (61), (65), (71) are explicitly stated for the first time in this paper.
- The third goal of the Bayesian description is to provide an intuitive graphical illustration that allows to easily overview a broad class of algorithms and to immediately identify their similarities and differences in terms of the underlying statistical assumptions.

By establishing new links between existing concepts, such an abstract overview should therefore also serve as a basis for revealing and exploring new directions. Note, however, that the review presented in this paper does not claim to cover all relevant acoustic model-based techniques and is rather meant as an inspiration to other researchers.

4 Uncertainty decoding

In the following, we consider the compensation rules of several uncertainty decoding techniques from a Bayesian view.

4.1 General example of uncertainty decoding

A fundamental example of uncertainty decoding can, e.g., be extracted from [1, 37–42]. The underlying observation model can be identified as

$$\mathbf{y}_n = \mathbf{x}_n + \mathbf{b}_n \text{ with } \mathbf{b}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\mathbf{b}_n}), \tag{10}$$

where $\mathbf{y}_n$ and $\mathbf{C}_{\mathbf{b}_n}$ often play the role of an enhanced feature vector, e.g., from a Wiener filtering front-end [40] and a measure of uncertainty from the enhancement process, respectively. Thus, the point estimate $\mathbf{y}_n$ can

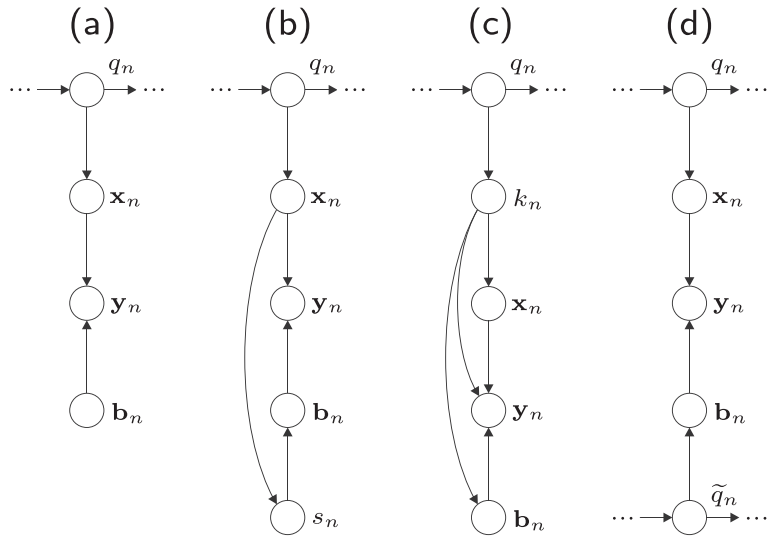

Fig. 2 Bayesian network representation of different model compensation techniques. Detailed descriptions of subfigures **a** to **d** are given in the text. Subfigure **d** is based on [4]

be seen as being enriched by the additional reliability information $\mathbf{C}_{\mathbf{b}_n}$. The observation model is representable by the Bayesian network in Fig. 2a. Exploiting the conditional independence properties of Bayesian networks [26], the compensation of the observation likelihood in (5) leads to [26]

$$\begin{aligned} p(\mathbf{y}_n|q_n) &= \int p(\mathbf{x}_n|q_n)\, p(\mathbf{y}_n|\mathbf{x}_n) d\mathbf{x}_n \\ &= \int \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_{\mathbf{x}|q_n}, \mathbf{C}_{\mathbf{x}|q_n})\, \mathcal{N}(\mathbf{y}_n; \mathbf{x}_n, \mathbf{C}_{\mathbf{b}_n}) d\mathbf{x}_n \\ &= \mathcal{N}(\mathbf{y}_n; \boldsymbol{\mu}_{\mathbf{x}|q_n}, \mathbf{C}_{\mathbf{x}|q_n} + \mathbf{C}_{\mathbf{b}_n}). \end{aligned} \tag{11}$$

Without loss of generality, a single Gaussian pdf $p(\mathbf{x}_n|q_n)$ is assumed since, in the case of a Gaussian mixture model (GMM), the linear mismatch function (10) can be applied to each Gaussian component separately.

4.2 Dynamic variance compensation

The concept of dynamic variance compensation [2] is based on a reformulation of the log-sum observation model [39]:

$$\mathbf{y}_n = \mathbf{x}_n + \log\left(1 + \exp(\widehat{\mathbf{r}}_n - \mathbf{x}_n)\right) + \mathbf{b}_n \tag{12}$$

with $\widehat{\mathbf{r}}_n$ being a noise estimate of any noise tracking algorithm and $\mathbf{b}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\mathbf{b}_n})$ a residual error term. Since the analytical derivation of $p(\mathbf{y}_n|q_n)$ is intractable, an approximate pdf is evaluated based on the assumption of $p(\mathbf{x}_n|\mathbf{y}_n)$ being Gaussian and that the compensation can be applied to each Gaussian component of the GMM separately [2]. According to Fig. 2a, the observation likelihood in (5), in its scaled version $\mathring{p}(\mathbf{y}_n|q_n)$, hence becomes:

$$\begin{aligned} \mathring{p}(\mathbf{y}_n|q_n) &= \int p(\mathbf{x}_n|q_n)\, \frac{p(\mathbf{x}_n|\mathbf{y}_n)}{p(\mathbf{x}_n)} d\mathbf{x}_n \\ &\approx \int p(\mathbf{x}_n|q_n)\, p(\mathbf{x}_n|\mathbf{y}_n) d\mathbf{x}_n \qquad (13) \\ &\approx \int \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_{\mathbf{x}|q_n}, \mathbf{C}_{\mathbf{x}|q_n}) \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}_n}, \mathbf{C}_{\mathbf{x}|\mathbf{y}_n}) d\mathbf{x}_n \\ &= \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}|q_n}; \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}_n}, \mathbf{C}_{\mathbf{x}|q_n} + \mathbf{C}_{\mathbf{x}|\mathbf{y}_n}), \qquad (14) \end{aligned}$$

where the approximation (13) can be justified if $p(\mathbf{x}_n)$ is assumed to be significantly "flatter," i.e., of larger variance, than $p(\mathbf{x}_n|\mathbf{y}_n)$. The estimation of the moments $\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}_n}, \mathbf{C}_{\mathbf{x}|\mathbf{y}_n}$ of $p(\mathbf{x}_n|\mathbf{y}_n)$ represents the core of [2].

4.3 Uncertainty decoding with SPLICE

The stereo piecewise linear compensation for environment (SPLICE) approach, first introduced in [43] and further developed in [44, 45], is a popular method for cepstral feature enhancement based on a mapping learned from stereo (i.e., clean and noisy) data [25]. While SPLICE can be used to derive an minimum mean square error (MMSE) [44] or MAP [43] estimate that is fed into the recognizer, it is also applicable in the context of uncertainty decoding [3], which we focus on in the following. In order to derive a Bayesian network representation of the uncertainty decoding version of SPLICE [3], we first note from [3] that one fundamental assumption is

$$p(\mathbf{x}_n|\mathbf{y}_n, s_n) = \mathcal{N}(\mathbf{x}_n; \mathbf{y}_n + \mathbf{r}_{s_n}, \boldsymbol{\Gamma}_{s_n}), \tag{15}$$

where s_n denotes a discrete region index, $\mathbf{r}_{s_n}$ is its bias, and $\mathbf{\Gamma}_{s_n}$ the uncertainty in that region. Exploiting the symmetry of the Gaussian pdf

$$\begin{aligned} p(\mathbf{x}_n|\mathbf{y}_n, s_n) &= \mathcal{N}(\mathbf{x}_n; \mathbf{y}_n + \mathbf{r}_{s_n}, \mathbf{\Gamma}_{s_n}) \\ &= \mathcal{N}(\mathbf{y}_n - \mathbf{x}_n; -\mathbf{r}_{s_n}, \mathbf{\Gamma}_{s_n}) \end{aligned} \tag{16}$$

and defining $\mathbf{b}_n = \mathbf{y}_n - \mathbf{x}_n$, we identify the observation model to be

$$\mathbf{y}_n = \mathbf{x}_n + \mathbf{b}_n \tag{17}$$

given a certain region index s_n. In the general case of s_n depending on $\mathbf{x}_n$, the observation model can be expressed by the Bayesian network in Fig. 2b with

$$p(\mathbf{b}_n|s_n) = \mathcal{N}(\mathbf{b}_n; -\mathbf{r}_{s_n}, \mathbf{\Gamma}_{s_n}). \tag{18}$$

This reveals that the introduction of different regions s_n is equivalent to assuming an affine model (18) with $p(\mathbf{b}_n)$ being a GMM instead of a single Gaussian density, as in (10). By introducing a separate prior model

$$\begin{aligned} p(\mathbf{y}_n) &= \sum_{s_n} p(s_n)\, p(\mathbf{y}_n|s_n) \\ &= \sum_{s_n} p(s_n)\, \mathcal{N}(\mathbf{y}_n; \boldsymbol{\mu}_{\mathbf{y}|s_n}, \mathbf{C}_{\mathbf{y}|s_n}), \end{aligned} \tag{19}$$

for the distorted speech $\mathbf{y}_n$, the likelihood in (5) can be adapted according to

$$\begin{aligned} p(\mathbf{y}_n|q_n) &= \int p(\mathbf{x}_n|q_n)\, p(\mathbf{y}_n|\mathbf{x}_n) d\mathbf{x}_n \\ &= \int p(\mathbf{x}_n|q_n)\, \frac{p(\mathbf{x}_n, \mathbf{y}_n)}{p(\mathbf{x}_n)} d\mathbf{x}_n = \int p(\mathbf{x}_n|q_n) \\ &\quad \cdot \frac{\sum_{s_n} p(\mathbf{x}_n|\mathbf{y}_n, s_n)\, p(\mathbf{y}_n|s_n)\, p(s_n)}{\sum_{s_n} \int p(\mathbf{x}_n|\mathbf{y}_n, s_n)\, p(\mathbf{y}_n|s_n)\, p(s_n) d\mathbf{y}_n} d\mathbf{x}_n. \end{aligned} \tag{20}$$

Although analytically tractable, both the numerator and the denominator in (20) are typically approximated for the sake of runtime efficiency [3].

4.4 Joint uncertainty decoding

Model-based joint uncertainty decoding [4] assumes an affine observation model in the cepstral domain

$$\mathbf{y}_n = \mathbf{A}_{k_n} \mathbf{x}_n + \mathbf{b}_n \tag{21}$$

with the deterministic matrix $\mathbf{A}_{k_n}$ and $p(\mathbf{b}_n|k_n) = \mathcal{N}(\mathbf{b}_n; \boldsymbol{\mu}_{\mathbf{b}|k_n}, \mathbf{C}_{\mathbf{b}|k_n})$ depending on the considered Gaussian component k_n of the GMM of the current HMM state q_n:

$$p(\mathbf{x}_n|q_n) = \sum_{k_n} p(k_n)\, p(\mathbf{x}_n|k_n). \tag{22}$$

The Bayesian network is depicted in Fig. 2c implying the following compensation rule:

$$p(\mathbf{y}_n|k_n) = \int p(\mathbf{x}_n|k_n)\, p(\mathbf{y}_n|\mathbf{x}_n, k_n) d\mathbf{x}_n, \tag{23}$$

which can be analytically derived analogously to (11). In practice, the compensation parameters $\mathbf{A}_{k_n}$, $\boldsymbol{\mu}_{\mathbf{b}|k_n}$, and $\mathbf{C}_{\mathbf{b}|k_n}$ are not estimated for each Gaussian component k_n but for each regression class comprising a set of Gaussian components [4].

4.5 REMOS

As many other techniques, the reverberation modeling for speech recognition (REMOS) concept [5, 46] assumes the environmental distortion to be additive in the melspectral domain. However, REMOS also considers the influence of the L previous clean speech feature vectors $\mathbf{x}_{n-L:n-1}$ in order to model the dispersive effect of reverberation and to relax the conditional independence assumption of conventional HMMs. The observation model reads in the logmelspec domain:

$$\begin{aligned} \mathbf{y}_n = \log \Bigg(& \exp(\mathbf{c}_n) + \exp(\mathbf{h}_n + \mathbf{x}_n) \\ & + \exp(\mathbf{a}_n) \odot \sum_{l=1}^{L} \exp(\boldsymbol{\mu}_l + \mathbf{x}_{n-l}) \Bigg), \end{aligned} \tag{24}$$

where the normally distributed random variables $\mathbf{c}_n, \mathbf{h}_n$, and $\mathbf{a}_n$ model the additive noise components, the early part of the room impulse response (RIR), and the weighting of the late part of the RIR, respectively, and the parameters $\boldsymbol{\mu}_{1:L}$ represent a deterministic description of the late part of the RIR. The Bayesian network is depicted in Fig. 3 with $\mathbf{b}_n = [\mathbf{c}_n, \mathbf{a}_n, \mathbf{h}_n]$. In contrast to most of the other compensation rules reviewed in this article, the REMOS concept necessitates a modification of the Viterbi decoder due to the introduced cross-connections in Fig. 3.

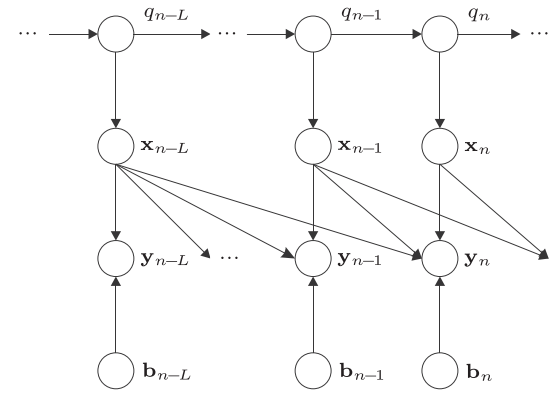

Fig. 3 Bayesian network representation of the REMOS concept (Subsection 4.5). The figure is based on [5]

In order to arrive at a computationally feasible decoder, the marginalization over the previous clean speech components $\mathbf{x}_{n-L:n-1}$ is circumvented by employing estimates $\widehat{\mathbf{x}}_{n-L:n-1}(q_{n-1})$ that depend on the best partial path, i.e., on the previous HMM state q_{n-1}. The resulting analytically intractable integral is then approximated by the maximum of its integrand:

$$\begin{aligned} &p(\mathbf{y}_n|q_n, \widehat{\mathbf{x}}_{n-L:n-1}(q_{n-1})) \\ &\quad = \int p(\mathbf{y}_n|\mathbf{x}_n, \widehat{\mathbf{x}}_{n-L:n-1}(q_{n-1}))\, p(\mathbf{x}_n|q_n) d\mathbf{x}_n \\ &\quad \approx \max_{\mathbf{x}_n} \; p(\mathbf{y}_n|\mathbf{x}_n, \widehat{\mathbf{x}}_{n-L:n-1}(q_{n-1}))\, p(\mathbf{x}_n|q_n). \end{aligned} \tag{25}$$

The determination of a global solution to (25) represents the core of the REMOS concept. The estimates $\widehat{\mathbf{x}}_{n-L:n-1}(q_{n-1})$ in turn are the solutions to (25) at previous time steps. We refer to [5] for a detailed derivation of the corresponding decoding routine.

It seems worthwhile noting that the simplification in (25) represents a variant of the MAP integral approximation, as often applied in Bayesian estimation [26]. To show this, we first omit the dependency on $\widehat{\mathbf{x}}_{n-L:n-1}(q_{n-1})$ for notational convenience and define

$$\begin{aligned} \mathbf{x}_n^{\text{MAP}} &= \arg\max_{\mathbf{x}_n} \; p(\mathbf{y}_n|\mathbf{x}_n)\, p(\mathbf{x}_n|q_n) \\ &= \arg\max_{\mathbf{x}_n} \; \frac{p(\mathbf{y}_n, \mathbf{x}_n|q_n)}{p(\mathbf{y}_n|q_n)} = \arg\max_{\mathbf{x}_n} \; p(\mathbf{x}_n|\mathbf{y}_n, q_n), \end{aligned} \tag{26}$$

where we scaled the objective function in the second step by the constant $1/p(\mathbf{y}_n|q_n)$. We can now reformulate the Bayesian integral leading to a novel derivation of (25)

$$\begin{aligned} p(\mathbf{y}_n|q_n) \;&=\; \int p(\mathbf{y}_n|\mathbf{x}_n)\, p(\mathbf{x}_n|q_n) d\mathbf{x}_n \\ &= \int p(\mathbf{y}_n|q_n)\, p(\mathbf{x}_n|\mathbf{y}_n, q_n) d\mathbf{x}_n \\ &\approx \int p(\mathbf{y}_n|q_n)\, p\left(\mathbf{x}_n|\mathbf{y}_n, q_n\right)\, \delta\left(\mathbf{x}_n - \mathbf{x}_n^{\text{MAP}}\right) d\mathbf{x}_n \\ &= p(\mathbf{y}_n|q_n)\, p\left(\mathbf{x}_n^{\text{MAP}}|\mathbf{y}_n, q_n\right) \qquad (27) \\ &= p\left(\mathbf{y}_n|\mathbf{x}_n^{\text{MAP}}\right)\, p\left(\mathbf{x}_n^{\text{MAP}}|q_n\right). \qquad (28) \end{aligned}$$

We see that the assumption underlying (25) is a modified MAP approximation:

$$p(\mathbf{x}_n|\mathbf{y}_n, q_n) \approx p(\mathbf{x}_n|\mathbf{y}_n, q_n)\, \delta\left(\mathbf{x}_n - \mathbf{x}_n^{\text{MAP}}\right), \tag{29}$$

which slightly differs from the conventional MAP approximation:

$$p(\mathbf{x}_n|\mathbf{y}_n, q_n) \approx \delta\left(\mathbf{x}_n - \mathbf{x}_n^{\text{MAP}}\right). \tag{30}$$

The obvious disadvantage of (29) is that the resulting score (25) does not represent a normalized likelihood w.r.t. to $\mathbf{y}_n$. On the other hand, the modified MAP approximation (29) leads to a scaled version of the exact likelihood $p(\mathbf{y}_n|q_n)$, cf. (27), with the scaling factor $p(\mathbf{x}_n^{\text{MAP}}|\mathbf{y}_n, q_n)$ being all the higher with increasing accuracy of the approximation (29).

4.6 Ion and Haeb-Umbach

Similarly to REMOS, the generic uncertainty decoding approach given in [24], and first proposed by [20], considers cross-connections in the Bayesian network in order to relax the conditional independence assumption of HMMs. The concept, as described in [24], is an example of uncertainty decoding, where the compensation rule can be defined by a modified Bayesian network structure—given in Fig. 4a—without fixing a particular functional form of the involved pdfs via an analytical observation model. In order to derive the compensation rule, we start by introducing the sequence $\mathbf{x}_{1:N}$ of latent clean speech vectors in each summand of (4)

$$\begin{aligned} p(\mathbf{y}_{1:N}, q_{1:N}) &= \int p(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}, q_{1:N}) d\mathbf{x}_{1:N} \\ &= \int p(\mathbf{y}_{1:N}|\mathbf{x}_{1:N}) \prod_{n=1}^{N} p(\mathbf{x}_n|q_n) p(q_n|q_{n-1}) d\mathbf{x}_{1:N} \\ &\sim \int \frac{p(\mathbf{x}_{1:N}|\mathbf{y}_{1:N})}{p(\mathbf{x}_{1:N})} \prod_{n=1}^{N} p(\mathbf{x}_n|q_n) p(q_n|q_{n-1}) d\mathbf{x}_{1:N}, \end{aligned} \tag{31}$$

where we exploited the conditional independence properties defined by Fig. 4a (respecting the dashed links) and dropped $p(\mathbf{y}_{1:N})$ in the last line of (31) as it represents a constant factor with respect to a varying state sequence $q_{1:N}$. The pdf in the numerator of (31) is next turned into

$$\begin{aligned} p(\mathbf{x}_{1:N}|\mathbf{y}_{1:N}) \;&=\; p(\mathbf{x}_1|\mathbf{y}_{1:N}) \prod_{n=2}^{N} p(\mathbf{x}_n|\mathbf{y}_{1:N}, \mathbf{x}_{1:n-1}) \\ &\approx \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{y}_{1:N}), \end{aligned} \tag{32}$$

where the conditional dependence (due to the head-to-head relation) of $\mathbf{x}_n$ and $\mathbf{x}_{1:n-1}$ is neglected. This corresponds to omitting the respective dashed links in Fig. 4a for each factor in (32) separately. The denominator in (31) can also be further decomposed if the dashed links in Fig. 4b, i.e., the head-to-tail relations in q_n, are disregarded:

$$p(\mathbf{x}_{1:N}) \approx \prod_{n=1}^{N} p(\mathbf{x}_n). \tag{33}$$

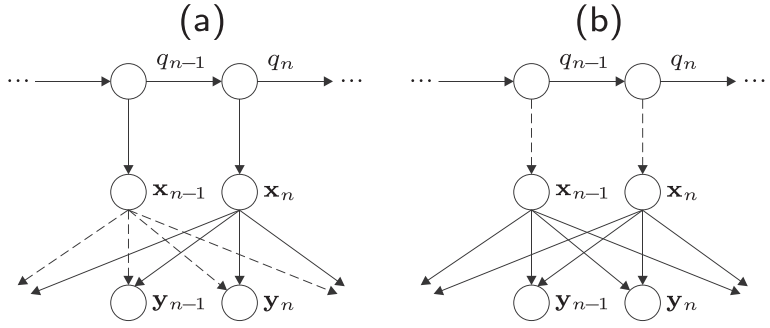

Fig. 4 Bayesian network representations **a** and **b** of the decoding rule of [24], where the *dashed links* are disregarded in the different steps of the derivation (Subsection 4.6)

With (32) and (33), the updated rule (31) is finally turned into the following simplified form:

$$p(\mathbf{y}_{1:N}, q_{1:N}) \sim \prod_{n=1}^{N} \int \frac{p(\mathbf{x}_n|\mathbf{y}_{1:N})}{p(\mathbf{x}_n)}\, p(\mathbf{x}_n|q_n) d\mathbf{x}_n\, p(q_n|q_{n-1}) \tag{34}$$

that is given in [24]. Due to the approximations in Fig. 4a, b, the compensation rule defined by (34) exhibits the same decoupling as (5) and can thus be carried out without modifying the underlying decoder. In practice, $p(\mathbf{x}_n)$ may, e.g., be modeled as a separate Gaussian density and $p(\mathbf{x}_n|\mathbf{y}_{1:N})$ as a separate Markov process [24].

4.7 Significance decoding

Assuming the affine model (10), the concept of significance decoding [9] first derives the moments of the posterior $p(\mathbf{x}_n|\mathbf{y}_n, q_n)$:

$$\begin{aligned} p(\mathbf{x}_n|\mathbf{y}_n, q_n) &= \frac{p(\mathbf{y}_n|\mathbf{x}_n, q_n)\, p(\mathbf{x}_n|q_n)}{\int p(\mathbf{y}_n|\mathbf{x}_n, q_n)\, p(\mathbf{x}_n|q_n) d\mathbf{x}_n} \\ &= \frac{p(\mathbf{y}_n|\mathbf{x}_n)\, p(\mathbf{x}_n|q_n)}{\int p(\mathbf{y}_n|\mathbf{x}_n)\, p(\mathbf{x}_n|q_n) d\mathbf{x}_n} \\ &= \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}_n,q_n}, \mathbf{C}_{\mathbf{x}|\mathbf{y}_n,q_n}), \end{aligned} \tag{35}$$

where the Bayesian network properties of Fig. 2a have been exploited in the numerator and the denominator and a single Gaussian pdf $p(\mathbf{x}_n|q_n)$ is assumed without loss of generality. In a second step, the clean likelihood $p(\mathbf{x}_n|q_n)$ is evaluated at $\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}_n,q_n}$ after adding the variance $\mathbf{C}_{\mathbf{x}|\mathbf{y}_n,q_n}$ to $\mathbf{C}_{\mathbf{x}|q_n}$, cf. (36).

In terms of probabilistic notation, this compensation rule corresponds to replacing the score calculation in (5) by an expected likelihood, similarly to [47, 48]:

$$\begin{aligned} p(\mathbf{y}_n|q_n) &\approx \mathcal{E}_{\mathbf{x}_n|\mathbf{y}_n,q_n}\{p(\mathbf{x}_n|q_n)\} \\ &= \int p(\mathbf{x}_n|\mathbf{y}_n, q_n)\, p(\mathbf{x}_n|q_n) d\mathbf{x}_n \\ &= \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}_n,q_n}; \boldsymbol{\mu}_{\mathbf{x}|q_n}, \mathbf{C}_{\mathbf{x}|\mathbf{y}_n,q_n} + \mathbf{C}_{\mathbf{x}|q_n}\right). \end{aligned} \tag{36}$$

For the case of single Gaussian densities, the (in the Bayesian sense) exact score $p(\mathbf{y}_n|q_n)$ is given by (11). Extending previous work [9], we show (11) to be bounded from above by the modified score (36):

$$\begin{aligned} \mathcal{E}_{\mathbf{x}_n|\mathbf{y}_n,q_n}\{p(\mathbf{x}_n|q_n)\} &= \int p(\mathbf{x}_n|\mathbf{y}_n, q_n)\, p(\mathbf{x}_n|q_n) d\mathbf{x}_n \\ &= \underbrace{\frac{\int p(\mathbf{y}_n|\mathbf{x}_n)\, p^2(\mathbf{x}_n|q_n) d\mathbf{x}_n}{p^2(\mathbf{y}_n|q_n)}}_{\alpha}\, p(\mathbf{y}_n|q_n), \end{aligned} \tag{37}$$

where α can be evaluated exploiting the product rules of Gaussians [26]:

$$\begin{aligned} \alpha &= \sqrt{\frac{\det(\mathbf{C}_{\mathbf{x}|q_n} + \mathbf{C}_{\mathbf{b}_n})}{\det(\mathbf{C}_{\mathbf{x}|q_n})}}\, \frac{\mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_{\mathbf{x}|q_n}, \frac{1}{2}\mathbf{C}_{\mathbf{x}|q_n} + \mathbf{C}_{\mathbf{b}_n})}{\mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_{\mathbf{x}|q_n}, \frac{1}{2}\mathbf{C}_{\mathbf{x}|q_n} + \frac{1}{2}\mathbf{C}_{\mathbf{b}_n})} \\ &\geq \sqrt{\frac{\det(\mathbf{C}_{\mathbf{x}|q_n} + \mathbf{C}_{\mathbf{b}_n})\, \det(\frac{1}{2}\mathbf{C}_{\mathbf{x}|q_n} + \frac{1}{2}\mathbf{C}_{\mathbf{b}_n})}{\det(\mathbf{C}_{\mathbf{x}|q_n})\, \det(\frac{1}{2}\mathbf{C}_{\mathbf{x}|q_n} + \mathbf{C}_{\mathbf{b}_n})}} \geq 1. \end{aligned} \tag{38}$$

A closer inspection of α reveals that the expected likelihood computation scales up $p(\mathbf{y}_n|q_n)$ for large values of $\mathbf{C}_{\mathbf{b}_n}$, which acts as an alleviation of the (potentially overly) flattening effect of $\mathbf{C}_{\mathbf{b}_n}$ on $p(\mathbf{y}_n|q_n)$, cf. (11).

5 Missing feature techniques

We next turn to missing feature techniques, which can be used to model feature distortion due to a front-end enhancement process [7], noise [49], or reverberation [50].

5.1 Feature vector imputation

A major subcategory of missing feature approaches is called *feature vector imputation* [6, 7, 27] where each feature vector component $y_n^{(d)}$, $d \in \{1,\ldots,D\}$, is either classified as reliable ($d \in \mathcal{R}_n$) or unreliable ($d \in \mathcal{U}_n$) with $\mathcal{R}_n$ and $\mathcal{U}_n$ denoting the set of reliable and unreliable components of the nth feature vector, respectively [24]. While unreliable components are withdrawn and replaced by an estimate $\widehat{x}_n^{(d)}$ of the original observation $y_n^{(d)}$, reliable components are directly "plugged" into the pdf. The score calculation in (5), in its scaled version $\mathring{p}(\mathbf{y}_n|q_n)$, therefore becomes

$$\begin{aligned}\mathring{p}(\mathbf{y}_n|q_n) &= \int p(\mathbf{x}_n|q_n)\,\frac{p(\mathbf{x}_n|\mathbf{y}_n)}{p(\mathbf{x}_n)}d\mathbf{x}_n \\ &\approx \int p(\mathbf{x}_n|q_n)\,p(\mathbf{x}_n|\mathbf{y}_n)d\mathbf{x}_n \end{aligned} \tag{39}$$

with

$$p(\mathbf{x}_n|\mathbf{y}_n) = \prod_{d=1}^{D} p(x_n^{(d)}|y_n^{(d)}) \tag{40}$$

and [24]

$$p(x_n^{(d)}|y_n^{(d)}) = \begin{cases} \delta(x_n^{(d)} - y_n^{(d)}) & d \in \mathcal{R} \\ \delta(x_n^{(d)} - \widehat{x}_n^{(d)}) & d \in \mathcal{U} \end{cases} \tag{41}$$

with the general Bayesian network in Fig. 5a. The approximation in (39) follows the same reasoning as (13).

5.2 Marginalization

The second major subcategory of missing feature techniques is called *marginalization* [6, 7, 27], where unreliable components are "replaced" by marginalizing over a clean-speech distribution $p(x_n^{(d)})$ that is usually not derived from the HMM but separately modeled. The posterior likelihood in (41) thus becomes [24]

$$p(x_n^{(d)}|y_n^{(d)}) = \begin{cases} \delta(x_n^{(d)} - y_n^{(d)}) & d \in \mathcal{R} \\ p(x_n^{(d)}) & d \in \mathcal{U} \end{cases} \tag{42}$$

with the general Bayesian network in Fig. 5a.

5.3 Modified imputation

The approach presented in [8] assumes the affine observation model (10) and evaluates the clean likelihood $p(\mathbf{x}_n|q_n)$ for an enhanced feature vector given by

$$\widehat{\mathbf{x}}_n = \arg\max_{\mathbf{x}_n}\; p(\mathbf{x}_n|q_n)\,p(\mathbf{y}_n|\mathbf{x}_n). \tag{43}$$

In order to fit this technique into our Bayesian perspective, we first consider $\widehat{\mathbf{x}}_n$ as a MAP estimate similar to (26). We furthermore note that for the affine model (10), we have

$$\begin{aligned} p(\mathbf{y}_n|\mathbf{x}_n) &= \mathcal{N}(\mathbf{y}_n;\mathbf{x}_n,\mathbf{C}_{\mathbf{b}_n}) \\ &= \mathcal{N}(\mathbf{x}_n;\mathbf{y}_n,\mathbf{C}_{\mathbf{b}_n}) = p(\mathbf{x}_n|\mathbf{y}_n). \end{aligned} \tag{44}$$

Starting off with the standard compensation rule following from (10), i.e., Fig. 2a, we show for the first time that [8] implicitly assumes $p(\mathbf{x}_n|\mathbf{y}_n)$ to be sharply peaked around the MAP estimate $\widehat{\mathbf{x}}_n$:

$$\begin{aligned} p(\mathbf{y}_n|q_n) &= \int p(\mathbf{x}_n|q_n)\,p(\mathbf{y}_n|\mathbf{x}_n)d\mathbf{x}_n \\ &= \int p(\mathbf{x}_n|q_n)\,p(\mathbf{x}_n|\mathbf{y}_n)d\mathbf{x}_n \\ &\approx \int p(\mathbf{x}_n|q_n)\,\delta(\mathbf{x}_n - \widehat{\mathbf{x}}_n)d\mathbf{x}_n, \end{aligned} \tag{45}$$

where (45) corresponds to evaluating the clean state-dependent likelihood at $\widehat{\mathbf{x}}_n$.

It seems finally interesting to note that (44) also explains the fact that the concept of modified imputation

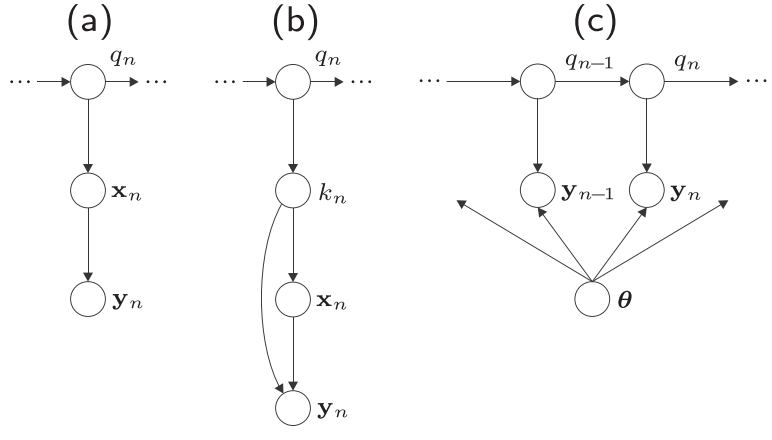

Fig. 5 Bayesian network representation of **a** different model compensation techniques, **b** CMLLR (Subsection 6.3) and MLLR (Subsection 6.4), and **c** MAP adaptation (Subsection 6.5). Detailed descriptions are given in the text

has been independently described in the following two forms:

$$\begin{aligned}\widehat{\mathbf{x}}_n &= \arg\max_{\mathbf{x}_n}\ p(\mathbf{x}_n|q_n)\,p(\mathbf{x}_n|\mathbf{y}_n)\\ &= \arg\max_{\mathbf{x}_n}\ p(\mathbf{x}_n|q_n)\,p(\mathbf{y}_n|\mathbf{x}_n)\end{aligned} \tag{46}$$

in [8] and [9], respectively.

6 Acoustic model adaptation and other model-based techniques

In the following, we consider the compensation rules of several acoustic model adaptation and other model-based approaches from a Bayesian view.

6.1 Parallel model combination

We start with the fundamental framework of parallel model combination (PMC) [10]. The observation model of the PMC concept is based on the log-sum distortion model and reads in the static logmelspec domain:

$$\mathbf{y}_n = \log\left(\alpha\exp(\mathbf{x}_n) + \exp(\mathbf{b}_n)\right), \tag{47}$$

where the deterministic parameter α accounts for level differences between the clean speech $\mathbf{x}_n$ and the distortion $\mathbf{b}_n$. Under the assumption of stationary distortions, i.e.,

$$p(\mathbf{b}_n) = \text{const.}, \tag{48}$$

the underlying Bayesian network corresponds to Fig. 2a. This explains the name of PMC as (47) combines two independent parallel models: the clean-speech HMM and the distortion model $p(\mathbf{b}_n)$. Since the resulting adapted pdf

$$p(\mathbf{y}_n|q_n) = \int p(\mathbf{x}_n|q_n)\,p(\mathbf{y}_n|\mathbf{x}_n,\mathbf{b}_n)\,p(\mathbf{b}_n)d(\mathbf{x}_n,\mathbf{b}_n) \tag{49}$$

cannot be derived in an analytical closed form, a variety of approximations to the true pdf $p(\mathbf{y}_n|q_n)$ have been investigated [10]. For nonstationary distortions, [10] proposes to employ a separate HMM for the distortion $\mathbf{b}_n$ leading to the Bayesian network representation of Fig. 2d. Marginalizing over the distortion state sequence $\widetilde{q}_{1:N}$ as in (5) reveals the acoustic score to become

$$\begin{aligned}p(\mathbf{y}_{1:N}|\mathbf{W}) &= \sum_{\substack{q_{1:N}\\ \widetilde{q}_{1:N}}} p(\mathbf{y}_{1:N}, q_{1:N}, \widetilde{q}_{1:N})\\ &= \sum_{\substack{q_{1:N}\\ \widetilde{q}_{1:N}}}\left\{\prod_{n=1}^{N} p(\mathbf{y}_n|q_n,\widetilde{q}_n)\,p(q_n|q_{n-1})\,p(\widetilde{q}_n|\widetilde{q}_{n-1})\right\},\end{aligned} \tag{50}$$

where

$$p(\mathbf{y}_n|q_n,\widetilde{q}_n) = \int p(\mathbf{x}_n|q_n)\,p(\mathbf{y}_n|\mathbf{x}_n,\mathbf{b}_n)\,p(\mathbf{b}_n|\widetilde{q}_n)d(\mathbf{x}_n,\mathbf{b}_n). \tag{51}$$

The overall acoustic score can be approximated by a 3D Viterbi decoder, which can in turn be mapped onto a conventional 2D Viterbi decoder [10].

6.2 Vector Taylor series model compensation

The concept of vector Taylor series (VTS) model compensation is frequently employed in practice yielding promising results [25]. Its fundamental idea is to linearize a nonlinear distortion model by a Taylor series [11, 51, 52]. The standard VTS approach [51] is based on the log-sum observation model:

$$\mathbf{y}_n = \log\left(\exp(\mathbf{h}_n + \mathbf{x}_n) + \exp(\mathbf{c}_n)\right), \tag{52}$$

where $p(\mathbf{h}_n) = \mathcal{N}(\mathbf{h}_n;\boldsymbol{\mu}_\mathbf{h},\mathbf{C}_\mathbf{h})$ captures short convolutive distortion and $p(\mathbf{c}_n) = \mathcal{N}(\mathbf{c}_n;\boldsymbol{\mu}_\mathbf{c},\mathbf{C}_\mathbf{c})$ models additive noise components. The Bayesian network is represented by Fig. 2a with $\mathbf{b}_n = [\,\mathbf{h}_n, \mathbf{c}_n]$. Note that in contrast to uncertainty decoding, $p(\mathbf{b}_n)$ is constant over time. As the adapted pdf is again of the form of (49) and thus analytically intractable, it is assumed that (52) can, firstly, be applied to each Gaussian component $p(\mathbf{x}_n|k_n)$ of the GMM

$$p(\mathbf{x}_n|q_n) = \sum_{k_n} p(k_n)p(\mathbf{x}_n|k_n) \tag{53}$$

individually and, secondly, be approximated by a Taylor series around $[\,\boldsymbol{\mu}_{\mathbf{x}|k_n},\boldsymbol{\mu}_\mathbf{h},\boldsymbol{\mu}_\mathbf{c}]$, where $\boldsymbol{\mu}_{\mathbf{x}|k_n}$ denotes the mean of the component $p(\mathbf{x}_n|k_n)$. There are various extensions to the VTS concept that are omitted here. For a more comprehensive review of VTS, we refer to [25].

6.3 CMLLR

Constrained maximum likelihood linear regression (CMLLR) [12, 53] can be seen as the deterministic counterpart of joint uncertainty decoding (Subsection 4.4) with the observation model

$$\mathbf{y}_n = \mathbf{A}_{k_n}\mathbf{x}_n + \mathbf{b}_{k_n} \tag{54}$$

and deterministic parameters $\mathbf{A}_{k_n},\mathbf{b}_{k_n}$. The adaptation rule of $p(\mathbf{y}_n|k_n)$ has the same form as (23) with

$$p(\mathbf{y}_n|\mathbf{x}_n,k_n) = \delta(\mathbf{y}_n - \mathbf{A}_{k_n}\mathbf{x}_n - \mathbf{b}_{k_n}). \tag{55}$$

With this reformulation, we identified the underlying Bayesian network to correspond to Fig. 5b, where the use of regression classes is again reflected by the dependency of the observation model parameters on the Gaussian component k_n (cf. Subsection 4.4). The affine observation model in (54) is equivalent to transforming the mean vector $\boldsymbol{\mu}_{\mathbf{x}|k_n}$ and covariance matrix $\mathbf{C}_{\mathbf{x}|k_n}$ of each Gaussian component of $p(\mathbf{x}_n|k_n)$:

$$\boldsymbol{\mu}_{\mathbf{y}|k_n} = \mathbf{A}_{k_n}\boldsymbol{\mu}_{\mathbf{x}|k_n} + \mathbf{b}_{k_n}, \tag{56}$$

$$\mathbf{C}_{\mathbf{y}|k_n} = \mathbf{A}_{k_n}\mathbf{C}_{\mathbf{x}|k_n}\mathbf{A}_{k_n}^T. \tag{57}$$

CMLLR represents a very popular adaptation technique due to its promising results and versatile fields of application, such as speaker adaptation [53], adaptive training [54] as well as noise [55] and reverberation-robust [56] ASR.

6.4 MLLR

The maximum likelihood linear regression (MLLR) concept [13] can be considered as a generalization of constrained maximum likelihood linear regression (CMLLR) as it allows for a separate transform matrix $\mathbf{B}_{k_n}$ in (57):

$$\boldsymbol{\mu}_{\mathbf{y}|k_n} = \mathbf{A}_{k_n}\boldsymbol{\mu}_{\mathbf{x}|k_n} + \mathbf{b}_{k_n}, \tag{58}$$

$$\mathbf{C}_{\mathbf{y}|k_n} = \mathbf{B}_{k_n}\mathbf{C}_{\mathbf{x}|k_n}\mathbf{B}_{k_n}^T. \tag{59}$$

In practice, however, MLLR is frequently applied to the mean vectors only [57–60] while neglecting the adaptation of the covariance matrix:

$$\mathbf{C}_{\mathbf{y}|k_n} = \mathbf{C}_{\mathbf{x}|k_n}. \tag{60}$$

This principle is also known from other approaches that are applicable to both means and variances but are often only carried out on the former (e.g., for the sake of robustness) [10, 61].

If applied to the mean vectors only, MLLR can in turn be considered as a simplified version of CMLLR, where the observation model (54) and the Bayesian network in Fig. 5b is assumed while the compensation of the variances is omitted.

The Bayesian network representation in Fig. 5b also underlies the general MLLR adaptation rule (58) and (59). In this case, however, it seems impossible to identify a corresponding observation model representation without analytically tying $\mathbf{A}_{k_n}$ and $\mathbf{B}_{k_n}$.

6.5 MAP adaptation

We next describe the MAP adaptation applied to any parameters $\boldsymbol{\theta}$ of the pdfs of an HMM and present a new formulation highlighting that these parameters are implicitly considered as Bayesian, i.e., random variables that are drawn once for all times as depicted in Fig. 5c [26]. As a direct consequence, any two observation vectors $\mathbf{y}_i, \mathbf{y}_j$ are conditionally dependent given the state sequence. The predictive pdf in (4) therefore explicitly depends on the adaptation data that we denote as $\mathbf{y}_{M:0}$, $M < 0$, and becomes

$$\begin{aligned} p(\mathbf{y}_{1:N}, q_{1:N}|\mathbf{y}_{M:0}) &= \int p(\mathbf{y}_{1:N}, q_{1:N}, \boldsymbol{\theta}|\mathbf{y}_{M:0})d\boldsymbol{\theta} \\ &= \int p(\mathbf{y}_{1:N}, q_{1:N}|\boldsymbol{\theta}, \mathbf{y}_{M:0})\, p(\boldsymbol{\theta}|\mathbf{y}_{M:0})d\boldsymbol{\theta} \\ &\approx p(\mathbf{y}_{1:N}, q_{1:N}|\boldsymbol{\theta}_{\mathrm{MAP}}, \mathbf{y}_{M:0}), \end{aligned} \tag{61}$$

where the posterior $p(\boldsymbol{\theta}|\mathbf{y}_{M:0})$ is approximated as Dirac distribution $\delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MAP}})$ at the mode $\boldsymbol{\theta}_{\mathrm{MAP}}$:

$$\begin{aligned} \boldsymbol{\theta}_{\mathrm{MAP}} &= \arg\max_{\boldsymbol{\theta}}\ p(\boldsymbol{\theta}|\mathbf{y}_{M:0}) \\ &= \arg\max_{\boldsymbol{\theta}}\ p(\mathbf{y}_{M:0}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}). \end{aligned} \tag{62}$$

An iterative (local) solution to (63) is obtained by the expectation maximization (EM) algorithm. Note that due to the MAP approximation of the posterior $p(\boldsymbol{\theta}|\mathbf{y}_{M:0})$, the conditional independence assumption is again fulfilled such that a conventional decoder can be employed.

6.6 Bayesian MLLR

As mentioned before, uncertainty decoding techniques allow for a time-varying pdf $p(\mathbf{b}_n)$, while model adaptation approaches, such as in Subsections 6.1, 6.2, and 6.6.1, mostly set $p(\mathbf{b}_n)$ to be constant over time. In both cases, however, the "randomized" model parameter $\mathbf{b}_n$ is assumed to be redrawn in each time step n as in Fig. 6a. In contrast, *Bayesian estimation*—as mentioned before—usually refers to inference problems, where the random model parameters are drawn once for all times [26] as in Fig. 6b.

Another example of Bayesian model adaptation, besides MAP, is Bayesian MLLR [14] applied to the mean vector $\boldsymbol{\mu}_{\mathbf{x}|q_n}$ of each pdf $p(\mathbf{x}_n|q_n)$:

$$\boldsymbol{\mu}_{\mathbf{y}|q_n} = \mathbf{A}\boldsymbol{\mu}_{\mathbf{x}|q_n} + \mathbf{c} \tag{63}$$

with $\mathbf{b} = [\mathbf{A}, \mathbf{c}]$ being usually drawn from a Gaussian distribution [14]. Here, we do not consider different regression classes and assume $p(\mathbf{x}_n|q_n)$ to be a single Gaussian pdf since, in the case of a GMM, the linear mismatch function (63) can be applied to each Gaussian component separately. The likelihood score in (4) thus becomes (in contrast to (61), we do not explicitly mention the dependency on the adaptation data $\mathbf{y}_{M:0}$ for notational convenience):

$$\begin{aligned} \sum_{q_{1:N}} p(\mathbf{y}_{1:N}, q_{1:N}) &= \sum_{q_{1:N}} \int p(\mathbf{y}_{1:N}, q_{1:N}, \mathbf{b})d\mathbf{b} \\ &= \sum_{q_{1:N}} \int p(\mathbf{y}_{1:N}, q_{1:N}|\mathbf{b})\, p(\mathbf{b})d\mathbf{b}. \end{aligned} \tag{64}$$

This score can, e.g., be approximated by a frame-synchronous Viterbi search [62]. Another approach is to apply the Bayesian integral in a frame-wise manner and use a conventional decoder [63]. In this case, we can establish an interesting link to the Bayesian network perspective by approximating the integral in (64) as follows:

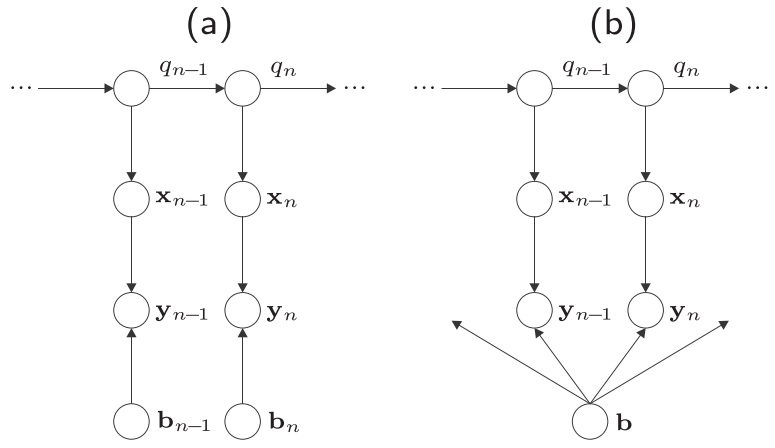

Fig. 6 Bayesian networks representing **a** typical uncertainty decoding and model adaptation with probabilistic parameter $\mathbf{b}_n$ and **b** Bayesian model adaptation

$$\begin{aligned}\int p(\mathbf{y}_{1:N}, q_{1:N}|\mathbf{b})\, p(\mathbf{b}) d\mathbf{b} \\ = \int \prod_{n=1}^{N} p(\mathbf{y}_n|q_n, \mathbf{b})\, p(q_n|q_{n-1})\, p(\mathbf{b}) d\mathbf{b} \\ \approx \prod_{n=1}^{N} \int p(\mathbf{y}_n|q_n, \mathbf{b})\, p(q_n|q_{n-1})\, p(\mathbf{b}) d\mathbf{b},\end{aligned} \tag{65}$$

where the original assumption of **b** being *identical* for all time steps n was relaxed to the case of **b** being *identically distributed* for all times steps n. The approximation in (65) can be interpreted as the conversion of the Bayesian network in Fig. 6b to the one in Fig. 6a with constant pdf $p(\mathbf{b}_n) = p(\mathbf{b})$ for all n.

6.6.1 Reverberant VTS

Reverberant VTS [15] is an extension of conventional VTS (Subsection 6.2) to capture the dispersive effect of reverberation. Its observation model reads for static features in the logmelspec domain:

$$\mathbf{y}_n = \log\left(\sum_{l=0}^{L} \exp(\mathbf{x}_{n-l} + \boldsymbol{\mu}_l) + \exp(\mathbf{b}_n)\right) \tag{66}$$

with $\mathbf{b}_n$ being an additive noise component modeled as normally distributed random variable and $\boldsymbol{\mu}_{0:L}$ being a deterministic description of the reverberant distortion. For the sake of tractability, the observation model is approximated in a similar manner as in the VTS approach. This concept can be seen as an alternative to REMOS (Subsection 4.5): While REMOS tailors the Viterbi decoder to the modified Bayesian network, reverberant VTS avoids the computationally expensive marginalization over all previous clean-speech vectors by averaging—and thus smoothing—the clean-speech statistics over all possible previous states and Gaussian components. Thus, $\mathbf{y}_n$ is assumed to depend on the extended clean-speech vector $\overline{\mathbf{x}}_n = [\,\mathbf{x}_{n-L}, \ldots, \mathbf{x}_n]$, cf. Fig. 7a vs. 7b.

6.7 Convolutive model adaptation

Besides the previously mentioned REMOS, reverberant VTS, and reverberant CMLLR concepts, there are three related approaches employing a convolutive observation model in order to describe the dispersive effect of reverberation [16–18]. All three approaches assume the following model in the logmelspec domain:

$$\mathbf{y}_n = \log\left(\sum_{l=0}^{L} \exp(\mathbf{x}_{n-l} + \boldsymbol{\mu}_l)\right), \tag{67}$$

where $\boldsymbol{\mu}_{0:L}$ denotes a deterministic description of the reverberant distortion that is differently determined by the three approaches. The observation model (67) can be represented by the Bayesian network in Fig. 3 without the random component $\mathbf{b}_n$. Both [16] and [18] use the "log-add approximation" [10] to derive $p(\mathbf{y}_n|q_n)$, i.e.,

$$\boldsymbol{\mu}_{\mathbf{y}|k_n} = \log\left(\exp(\boldsymbol{\mu}_{\mathbf{x}|k_n} + \boldsymbol{\mu}_0) + \sum_{l=1}^{L} \exp(\overline{\boldsymbol{\mu}}_{\mathbf{x}|q_{n-l}} + \boldsymbol{\mu}_l)\right), \tag{68}$$

where $\boldsymbol{\mu}_{\mathbf{y}|k_n}$ and $\boldsymbol{\mu}_{\mathbf{x}|k_n}$ denote the mean of the k_n-th Gaussian component of $p(\mathbf{y}_n|q_n)$ and $p(\mathbf{x}_n|q_n)$, respectively. The previous means $\overline{\boldsymbol{\mu}}_{\mathbf{x}|q_{n-l}}$, $l > 0$ are averaged over all means of the corresponding GMM $p(\mathbf{x}_{n-l}|q_{n-l})$. On the other hand, [17] employs the "log-normal approximation" [10] to adapt $p(\mathbf{y}_n|q_n)$ according to (67). While [16] and [17] perform the adaptation once prior to recognition and then use a standard decoder, the concept proposed in [18]

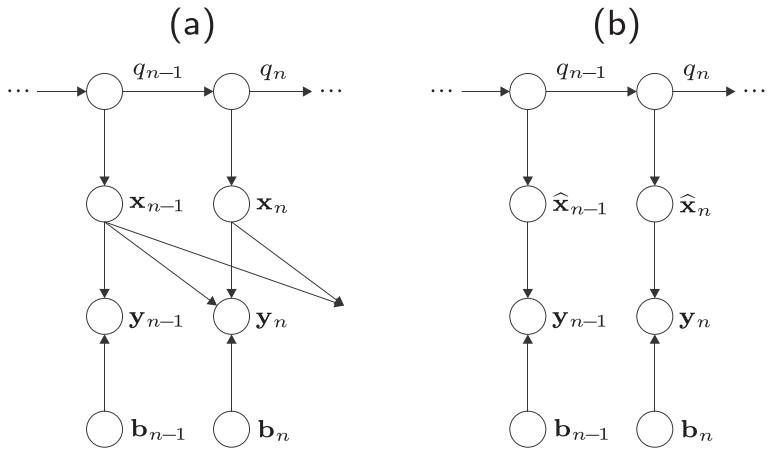

Fig. 7 Bayesian network representation of reverberant VTS (Subsection 6.6.1) **a** before and **b** after approximation via an extended observation vector. The figure is based on [15]

performs an online adaptation based on the best partial path [15].

It should be pointed out here that there is a variety of other approximations to the statistics of the log-sum of (mixtures of) Gaussian random variables (as seen in Subsections 4.2, 4.5, 6.1, 6.2, 6.6.1), ranging from different PMC methods [10] to maximum [64], piecewise linear [65], and other analytical approximations [66–70].

6.8 Takiguchi et al.

In contrast to the approaches of Subsections 6.6.1 and 6.7, the concept proposed in [19] assumes the reverberant observation vector $\mathbf{y}_{n-1}$ at time $n-1$ to be an approximation to the reverberation tail at time n in the logmelspec domain:

$$\mathbf{y}_n = \log\left(\exp(\mathbf{h}+\mathbf{x}_n) + \exp(\boldsymbol{\alpha}+\mathbf{y}_{n-1})\right), \tag{69}$$

where $\mathbf{h}$ and $\boldsymbol{\alpha}$ are deterministic parameters modeling short convolutive distortion and the weighting of the reverberation tail, respectively. We link this approach to the Bayesian framework by rewriting each summand in (4) according to (69):

$$p(\mathbf{y}_{1:N}, q_{1:N}) = \prod_{n=1}^{N} p(\mathbf{y}_n|q_n, \mathbf{y}_{n-1})\, p(q_n|q_{n-1}) \tag{70}$$

with the Bayesian network of Fig. 8. It seems interesting to note that (69) can be analytically evaluated as $\mathbf{y}_{n-1}$ is observed and, thus, (69) represents a nonlinear mapping $g(\cdot)$ of one random vector $\mathbf{x}_n$: $\mathbf{y}_n = g(\mathbf{x}_n)$ with

$$p(\mathbf{y}_n|q_n, \mathbf{y}_{n-1}) = \frac{p(\mathbf{x}_n|q_n)}{\det(J_{\mathbf{y}_n}(g^{-1}(\mathbf{y}_n))}, \tag{71}$$

where $\mathbf{x}_n = g^{-1}(\mathbf{y}_n)$ and $J_{\mathbf{y}_n}$ denotes the Jacobian w.r.t. $\mathbf{y}_n$.

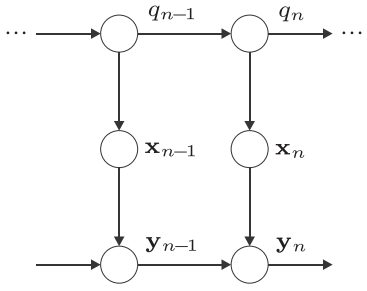

Fig. 8 Bayesian network representation of [19] (Subsection 6.8)

6.9 Conditional HMMs [22] and combined-order HMMs [23]

We close this section by broadening the view and pointing to two model-based approaches that cannot be classified as "model adaptation" as they postulate different HMM topologies rather than adapting a conventional HMM. Both approaches aim at relaxing the conditional independence assumption of conventional HMMs in order to improve the modeling of the inter-frame correlation.

The concept of conditional HMMs [22] models the observation $\mathbf{y}_n$ as depending on the previous observations at time shifts $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_P) \in \mathbb{N}^P$. Each summand in (4) therefore becomes

$$p(\mathbf{y}_{1:N}, q_{1:N}) = \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{y}_{n-\psi_1}, \ldots, \mathbf{y}_{n-\psi_P}, q_n)\, p(q_n|q_{n-1}) \tag{72}$$

according to Fig. 9a. Such HMMs are also known as *autoregressive HMMs* [26].

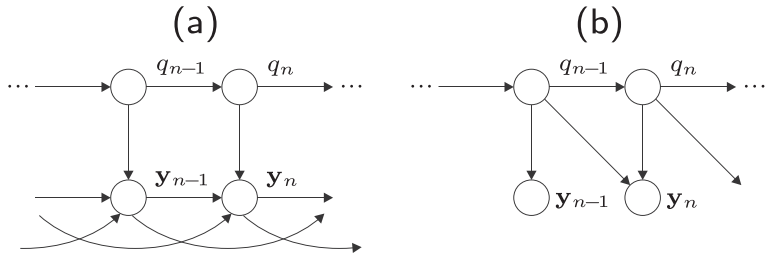

Fig. 9 Bayesian network representation of **a** conditional and **b** combined-order HMMS (Subsection 6.9). Subfigure **a** is based on [26]

In contrast to conditional HMMs, combined-order HMMs [23] assume the current observation $\mathbf{y}_n$ to depend on the previous HMM state q_{n-1} in addition to state q_n:

$$p(\mathbf{y}_{1:N}, q_{1:N}) = \prod_{n=1}^{N} p(\mathbf{y}_n|q_n, q_{n-1})\, p(q_n|q_{n-1}) \tag{73}$$

according to Fig. 9b, which can be thought of as a conventional first-order HMM with a second output pdf per state.

While conditional HMMs represent the statistically more accurate model for correlated speech feature vectors, combined-order HMMs circumvent the mathematically more complex inference step by a larger number of HMM parameters [23].

7 Relevance for DNN-based ASR

Before concluding this article, we build the bridge of the discussed model-based techniques for GMM-HMM-based ASR systems to the recent deep learning-based architectures.

The most immediate approach of exploiting conventional model-based techniques is within the framework of bottleneck or tandem systems [25]. There, deep neural networks (DNNs) are used for feature extraction while the ASR system's acoustic model is based on GMM-HMMs. For such systems, the presented approaches could—in principal—be applied in the same way as for conventional GMM-HMMs. However, the definition of meaningful observation models seems less intuitive as the features undergo various nonlinear transforms before being presented to the GMM-HMM system.

A popular alternative to bottleneck and tandem systems are the DNN-HMM hybrid approaches [71], which are often used along with logmelspec features (and their derivatives) as input domain [72, 73]. While—from a Bayesian perspective—the network representations of GMM-HMM and DNN-HMM systems are the same, cf. Fig. 1a, their decisive difference is the definition of the state-dependent output pdfs, which read in the case of DNN-HMMs:

$$p(\mathbf{x}_n|q_n) = \frac{p(q_n|\mathbf{x}_n)\, p(\mathbf{x}_n)}{p(q_n)} \sim \frac{p(q_n|\mathbf{x}_n)}{p(q_n)} \tag{74}$$

with $p(q_n|\mathbf{x}_n)$ being the q_nth output node of the DNN, $p(q_n)$ being the prior probability of each HMM state (senone), estimated from the training set, and $p(\mathbf{x}_n)$ being independent of the word sequence and thus to be ignored [71].

There are various approaches for adapting a DNN-HMM to changing acoustic environments or speaker characteristics. Most of them aim at either adapting certain weights of the DNN itself, such as in [74–76], or at presenting transformed/enriched input features to the DNN, as in [72, 77]. In the following, we will briefly discuss the application of the Bayesian perspective, taken on in this article, to DNN-HMMs. Analogously to the previous sections, we start with the definition of an exemplary observation model:

$$\mathbf{x}_n = \mathbf{y}_n + \mathbf{b}_n, \tag{75}$$

where $\mathbf{b}_n$ can again be a deterministic or random variable and, in contrast to the previous sections, the observation model is resolved for $\mathbf{x}_n$. With (75) and Fig. 1b, the adaptation of the q_nth output node of a DNN yields:

$$\begin{aligned} p(q_n|\mathbf{y}_n) &= \frac{\int p(\mathbf{y}_n, \mathbf{x}_n|q_n) d\mathbf{x}_n\, p(q_n)}{p(\mathbf{y}_n)} \\ &= \frac{\int p(\mathbf{x}_n|q_n)\, p(\mathbf{y}_n|\mathbf{x}_n, q_n) d\mathbf{x}_n\, p(q_n)}{p(\mathbf{y}_n)} \\ &= \frac{\int \frac{p(q_n|\mathbf{x}_n)p(\mathbf{x}_n)}{p(q_n)}\, p(\mathbf{y}_n|\mathbf{x}_n) d\mathbf{x}_n\, p(q_n)}{p(\mathbf{y}_n)} \\ &= \int p(q_n|\mathbf{x}_n)\, p(\mathbf{x}_n|\mathbf{y}_n) d\mathbf{x}_n, \end{aligned} \tag{76}$$

where $p(\mathbf{x}_n|\mathbf{y}_n)$ is defined through (75). In theory, any of the previously discussed observation models could thus be directly applied to a DNN-HMM as long as resolving them for $\mathbf{x}_n$ is feasible. In practice, however, both the parameter estimation step as well as the compensation step (76) can become complex.

In case of $\mathbf{b}_n$ being a random variable, the solution of (76) yields an interesting interpretation, which can be revealed by considering the q_n-th output node of the DNN as a transform $f(\cdot)$ of $\mathbf{x}_n$,

$$p(q_n|\mathbf{x}_n) = f_{q_n}(\mathbf{x}_n) = z_{q_n}, \tag{77}$$

and exploiting the fundamental property of variable transform in expectation operators:

$$\begin{aligned} p(q_n|\mathbf{y}_n) &= \int f_{q_n}(\mathbf{x}_n)\, p(\mathbf{x}_n|\mathbf{y}_n) d\mathbf{x}_n = \mathcal{E}\{f_{q_n}(\mathbf{x}_n)|\mathbf{y}_n\} \\ &= \mathcal{E}\{z_{q_n}|\mathbf{y}_n\} = \int z_{q_n}\, p(z_{q_n}|\mathbf{y}_n) dz_{q_n}. \end{aligned} \tag{78}$$

In other words, the DNN adaptation (76) corresponds to deriving the mean of the transformed random variable

$$z_{q_n} = f_{q_n}(\mathbf{x}_n) \overset{(75)}{=} f_{q_n}(\mathbf{y}_n + \mathbf{b}_n) \tag{79}$$

given the observation $\mathbf{y}_n$, which can, e.g., be achieved by numerical or deterministic integral approximations [78].

In case of $\mathbf{b}_n$ being a deterministic parameter, (76) simplifies to evaluating the DNN for the transformed observation:

$$\begin{aligned} p(q_n|\mathbf{y}_n) &= \int p(q_n|\mathbf{x}_n)\, p(\mathbf{x}_n|\mathbf{y}_n) d\mathbf{x}_n \\ &= \int p(q_n|\mathbf{x}_n)\, \delta(\widehat{\mathbf{x}}_n - \mathbf{x}_n) d\mathbf{x}_n \\ &= f_{q_n}(\widehat{\mathbf{x}}_n) \overset{(75)}{=} f_{q_n}(\mathbf{y}_n + \mathbf{b}_n). \end{aligned} \tag{80}$$

If the transform parameters (here: $\mathbf{b}_n$) are estimated irrespectively of the ASR system's acoustic model, (80) can be seen as a "conventional" feature enhancement step. If the transform parameters are discriminatively estimated using error back-propagation through the DNN, (80) could also be considered as adaptation of the DNN's input layer weights.

8 Conclusions

In this article, we described the compensation rules of several acoustic model-based techniques employing the Bayesian formalism. Some of the presented Bayesian descriptions are already given in the original papers and others can be easily derived based on the original papers (cf. Subsections 4.3, 4.4, 4.6, and 6.9). Beyond this, however, the links of the decoding rules of the concepts of REMOS (Subsection 4.5), significance decoding (Subsection 4.7), modified imputation (Subsection 5.3), CMLLR/MLLR (Subsections 6.3 and 6.4), MAP (Subsection 6.5), Bayesian MLLR (Subsection 6.6), and Takiguchi et al. [19] (Subsection 6.8) to the Bayesian framework via the mathematical reformulations in (28), (37), (45), (55), (61), (65), and (71), respectively, are explicitly stated for the first time in this paper.

As a byproduct of the Bayesian formalism, the considered concepts are represented here as Bayesian networks, which both highlights and hides certain crucial aspects. Most importantly, neither the particular functional form of the joint pdf nor potential approximations to arrive at a tractable algorithm nor the provenance of (i.e., the estimation procedure for) the compensation parameters are reflected.

On the other hand, the Bayesian network description provides a convenient representation to immediately and clearly identify some major properties:

- The cross-connections depicted in Figs. 3, 4, 7a, 8, and 9 show that the underlying concept aims at improving the modeling of the inter-frame correlation, e.g., to increase the robustness of the acoustic model against reverberation. If applied in a straightforward way, such cross-connections would entail a costly modification of the Viterbi decoder. In this paper, we summarized some important approximations that allow for a more efficient decoding of the extended Bayesian network, cf. Subsections 4.5, 4.6, 6.6.1, and 6.7. Some of these typically empirically motivated or just intuitive approximations, especially neglected statistical dependencies, become obvious from a Bayesian network, as shown in Figs. 4 and 7.
- The approaches introducing instantaneous (here: purely vertical) extensions to the Bayesian network, as in Figs. 2a–c and 5c, usually aim at compensating for nondispersive distortions, such as additive or short-ranging convolutive noise.
- The arcs in Figs. 2c and 5b illustrate that the observed vector $\mathbf{y}_n$ does not only depend on the state q_n (or mixture component k_n) through $\mathbf{x}_n$. As a consequence, one can deduce that the compensation parameters do depend on the phonetic content, as in Subsections 4.4, 6.3, and 6.4.
- The graphical model representation also succinctly highlights whether a Bayesian modeling paradigm is applied, as in Figs. 5c and 6b, or not, as in Figs. 5a, b.
- The existence of the additional latent variable $\mathbf{x}_n$ in most of the presented Bayesian network representations expresses that an explicit observation model or an implicit statistical model between the clean and the corrupted features is employed. In contrast, the graphical representations in Figs. 5c and 9 show that—instead of a distinct compensation step—a modified HMM topology is used.

In summary, the condensed description of the various concepts from the same Bayesian perspective shall allow other researchers to more easily exploit or combine existing techniques and to relate their own algorithms to the presented ones. This seems all the more important as

the recent acoustic modeling approaches based on DNNs raise new challenges for the conventional robustness techniques [25].

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
The authors would like to thank the Deutsche Forschungsgemeinschaft (DFG) for supporting this work (contract number KE 890/4-2).

Author details
[1]Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg, Cauerstr. 7, 91058 Erlangen, Germany. [2]Faculty of Electrical Engineering and Information Technology, Ostbayerische Technische Hochschule Regensburg, Seybothstr. 2, 93053 Regensburg, Germany.

Received: 15 April 2015 Accepted: 12 November 2015
Published online: 02 December 2015

References

1. JA Arrowood, MA Clements, in *Proc. ICSLP*. Using Observation Uncertainty in HMM Decoding (ISCA, Baixas, France, 2002), pp. 1561–1564
2. L Deng, J Droppo, A Acero, Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. IEEE Trans. Speech Audio Process. **13**(3), 412–421 (2005)
3. J Droppo, A Acero, L Deng, in *Proc. ICASSP*. Uncertainty Decoding with SPLICE for Noise Robust Speech Recognition, vol. 1 (IEEE, New Jersey, USA, 2002), pp. 57–60
4. H Liao, *Uncertainty decoding for noise robust speech recognition, PhD thesis*. (Univ. of Cambridge, 2007). http://mi.eng.cam.ac.uk/~mjfg/thesis_hl251.pdf
5. R Maas, W Kellermann, A Sehr, T Yoshioka, M Delcroix, K Kinoshita, T Nakatani, in *Proc. Int. Conf. on Digital Signal Process*. Formulation of the REMOS Concept from an Uncertainty Decoding Perspective (IEEE, New Jersey, USA, 2013)
6. M Cooke, P Green, L Josifovski, A Vizinho, Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data. Speech Commun. **34**(3), 267–285 (2001)
7. B Raj, RM Stern, Missing-feature approaches in speech recognition. IEEE Signal Process. Mag. **22**(5), 101–116 (2005)
8. D Kolossa, A Klimas, R Orglmeister, in *Proc. WASPAA*. Separation and Robust Recognition of Noisy, Convolutive Speech Mixtures Using Time-Frequency Masking and Missing Data Techniques (IEEE, New Jersey, USA, 2005), pp. 82–85
9. AH Abdelaziz, D Kolossa, in *Proc. Interspeech*. Decoding of Uncertain Features Using the Posterior Distribution of the Clean Data for Robust Speech Recognition (ISCA, Baixas, France, 2012)
10. MJF Gales, Model-based techniques for noise robust speech recognition, PhD thesis (1995). http://mi.eng.cam.ac.uk/~mjfg/thesis.pdf
11. A Acero, L Deng, T Kristjansson, J Zhang, in *Proc. ICSLP*. HMM Adaptation Using Vector Taylor Series for Noisy Speech Recognition, vol. 3 (ISCA, Baixas, France, 2000), pp. 869–872
12. VV Digalakis, D Rtischev, LG Neumeyer, Speaker adaptation using constrained estimation of Gaussian mixtures. IEEE Trans. Speech Audio Process. **3**(5), 357–366 (1995)
13. CJ Leggetter, PC Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput. Speech Lang. **9**(2), 171–185 (1995)
14. J-T Chien, Linear regression based Bayesian predictive classification for speech recognition. IEEE Trans. Speech Audio Process. **11**(1), 70–79 (2003)
15. YQ Wang, MJF Gales, in *Proc. ASRU*. Improving Reverberant VTS for Hands-Free Robust Speech Recognition (IEEE, New Jersey, USA, 2011), pp. 113–118
16. HG Hirsch, H Finster, A new approach for the adaptation of HMMs to reverberation and background noise. Speech Commun. **50**(3), 244–263 (2008)
17. CK Raut, T Nishimoto, S Sagayama, in *Proc. ICASSP*. Model Adaptation for Long Convolutional Distortion by Maximum Likelihood Based State Filtering Approach, vol. 1 (IEEE, New Jersey, USA, 2006), pp. 1133–1136
18. A Sehr, R Maas, W Kellermann, in *Proc. ICASSP*. Frame-wise HMM Adaptation Using State-Dependent Reverberation Estimates (IEEE, New Jersey, USA, 2011), pp. 5484–5487
19. T Takiguchi, M Nishimura, Y Ariki, Acoustic model adaptation using first-order linear prediction for reverberant speech. IEICE Trans. Inform. Syst. **E89-D**(3), 908–914 (2006)
20. V Ion, R Haeb-Umbach, A novel uncertainty decoding rule with applications to transmission error robust speech recognition. IEEE Trans. Audio, Speech, Lang. Process. **16**(5), 1047–1060 (2008)
21. JL Gauvain, CH Lee, in *Proc. Workshop on Speech and Natural Lang*. MAP Estimation of Continuous Density HMM: Theory and Applications (Morgan Kaufmann, Burlington, USA, 1992), pp. 185–190
22. J Ming, FJ Smith, Modelling of the interframe dependence in an HMM using conditional Gaussian mixtures. Comput. Speech Lang. **10**, 229–247 (1996)
23. R Maas, SR Kotha, A Sehr, W Kellermann, in *Proc. Int. Workshop on Cognitive Inform. Process*. Combined-Order Hidden Markov Models for Reverberation-Robust Speech Recognition (IEEE, New Jersey, USA, 2012), pp. 167–171
24. R Haeb-Umbach, in *Robust Speech Recognition of Uncertain or Missing Data*, ed. by D Kolossa, R Haeb-Umbach. Uncertainty Decoding and Conditional Bayesian Estimation (Springer, Berlin Heidelberg, 2011), pp. 9–33
25. J Li, L Deng, Y Gong, R Haeb-Umbach, An overview of noise-robust automatic speech recognition. IEEE/ACM Trans. Audio, Speech, Lang. Process. **22**(4), 745–777 (2014)
26. CM Bishop, *Pattern Recognition and Machine Learning*. (Springer, New York, 2006)
27. D Kolossa, R Haeb-Umbach, *Robust Speech Recognition of Uncertain or Missing Data*. (Springer, Berlin Heidelberg, 2011)
28. G Heigold, H Ney, R Schlüter, S Wiesler, Discriminative training for automatic speech recognition: modeling, criteria, optimization, implementation, and performance. IEEE Signal Process. Mag. **29**(6), 58–69 (2012)
29. M Matassoni, M Omologo, D Giuliani, P Svaizer, Hidden Markov model training with contaminated speech material for distant-talking speech recognition. Comput. Speech Lang. **16**(2), 205–223 (2002)
30. A Sehr, C Hofmann, R Maas, W Kellermann, in *Proc. Interspeech*. A Novel Approach for Matched Reverberant Training of HMMs Using Data Pairs (ISCA, Baixas, France, 2010), pp. 566–569
31. Y Gong, Speech recognition in noisy environments: A survey. Speech Commun. **16**(3), 261–291 (1995). doi:10.1016/0167-6393(94)00059-J
32. C-H Lee, On stochastic feature and model compensation approaches to robust speech recognition. Speech Commun. **25**(1–3), 29–47 (1998). doi:10.1016/S0167-6393(98)00028-4
33. Q Huo, C-H Lee, Robust speech recognition based on adaptive classification and decision strategies. Speech Commun. **34**(1–2), 175–194 (2001). doi:10.1016/S0167-6393(00)00053-4
34. J Droppo, in *Springer Handbook of Speech Processing*. Environmental Robustness (Springer, Berlin Heidelberg, 2008), pp. 653–680
35. T Yoshioka, A Sehr, M Delcroix, K Kinoshita, R Maas, T Nakatani, W Kellermann, Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition. IEEE Signal Process. Mag. **29**(6), 114–126 (2012)
36. T Virtanen, R Singh, B Raj, *Techniques for Noise Robustness in Automatic Speech Recognition*. (Wiley, UK, 2013)
37. JN Holmes, WJ Holmes, PN Garner, in *Proc. Eurospeech*. Using Formant Frequencies in Speech Recognition, vol. 97 (ISCA, Baixas, France, 1997), pp. 2083–2087
38. TT Kristjansson, BJ Frey, in *Proc. ICASSP*. Accounting for Uncertainity in Observations: A New Paradigm for Robust Speech Recognition, vol. 1 (IEEE, New Jersey, USA, 2002), pp. 61–64
39. L Deng, J Droppo, A Acero, in *Proc. ICSLP*. Exploiting Variances in Robust Feature Extraction Based on a Parametric Model of Speech Distortion, vol. 4, (2002), pp. 2449–2452
40. MC Benitez, JC Segura, A Torre, J Ramirez, A Rubio, in *Proc. ICSLP*. Including Uncertainty of Speech Observations in Robust Speech Recognition (ISCA, Baixas, France, 2004), pp. 137–140

41. V Stouten, H Van Hamme, P Wambacq, Model-based feature enhancement with uncertainty decoding for noise robust ASR. Speech Commun. **48**(11), 1502–1514 (2006)
42. M Delcroix, T Nakatani, S Watanabe, Static and dynamic variance compensation for recognition of reverberant speech with dereverberation preprocessing. IEEE Trans. Audio, Speech, Lang. Process. **17**(2), 324–334 (2009)
43. L Deng, A Acero, M Plumpe, XD Huang, in *Proc. ICSLP*. Large Vocabulary Continuous Speech Recognition Under Adverse Conditions, vol. 3 (ISCA, Baixas, France, 2000), pp. 806–809
44. L Deng, A Acero, L Jiang, J Droppo, X Huang, in *Proc. ICASSP*. High-Performance Robust Speech Recognition Using Stereo Training Data, vol. 1 (IEEE, New Jersey, USA, 2001), pp. 301–304
45. L Deng, J Droppo, A Acero, Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition. IEEE Trans. Speech Audio Process. **11**(6), 568–580 (2003)
46. A Sehr, R Maas, W Kellermann, Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition. IEEE Trans. Audio, Speech, Lang. Process. **18**(7), 1676–1691 (2010)
47. JA Arrowood, *Using observation uncertainty for robust speech recognition, PhD thesis*. (Georgia Institute of Technology, 2003). https://smartech.gatech.edu/bitstream/handle/1853/5383/arrowood_jon_a_200312_phd.pdf
48. RF Astudillo, R Orglmeister, Computing MMSE estimates and residual uncertainty directly in the feature domain of ASR using STFT domain speech distortion models. IEEE Trans. Audio, Speech, Lang. Process. **21**(5), 1023–1034 (2013)
49. M Cooke, A Morris, P Green, in *Proc. ICASSP*. Missing Data Techniques for Robust Speech Recognition, vol. 2, (1997), pp. 863–866
50. KJ Palomäki, GJ Brown, JP Barker, Techniques for handling convolutional distortion with 'missing data' automatic speech recognition. Speech Commun. **43**(1–2), 123–142 (2004)
51. PJ Moreno, *Speech recognition in noisy environments, PhD thesis*. (Carnegie Mellon Univ., Pittsburgh, 1996). http://www.cs.cmu.edu/~robust/Thesis/pjm_thesis.pdf
52. J Li, L Deng, D Yu, Y Gong, A Acero, A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions. Comput. Speech Lang. **23**(3), 389–405 (2009)
53. MJF Gales, Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. **12**, 75–98 (1997)
54. K Yu, MJF Gales, Discriminative cluster adaptive training. IEEE Trans. Audio, Speech, Lang. Process. **14**(5), 1694–1703 (2006)
55. E Vincent, J Barker, S Watanabe, J Le Roux, F Nesta, M Matassoni, in *Proc. ASRU*. The Second 'CHiME' Speech Separation and Recognition Challenge: An Overview of Challenge Systems and Outcomes (IEEE, New Jersey, USA, 2013), pp. 162–167
56. K Kinoshita, M Delcroix, T Yoshioka, T Nakatani, A Sehr, W Kellermann, R Maas, in *Proc. WASPAA*. The REVERB Challenge: A common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech, (2013)
57. T Anastasakos, J McDonough, R Schwartz, J Makhoul, in *Proc. ICSLP*. A Compact Model for Speaker-Adaptive Training, vol. 2 (ISCA, Baixas, France, 1996), pp. 1137–1140
58. S Young, G Evermann, D Kershaw, G Moore, J Odell, D Ollason, D Povey, V Valtchev, P Woodland, *The HTK Book*. (Cambridge Univ. Eng. Dept., UK, 2002)
59. M Delcroix, K Kinoshita, T Nakatani, S Araki, A Ogawa, T Hori, S Watanabe, M Fujimoto, T Yoshioka, T Oba, *et al*, in *Proc. Int. Workshop on Mach. Listening in Multisource Environments (CHiME)*. Speech Recognition in the Presence of Highly Non-stationary Noise Based on Spatial, Spectral and Temporal Speech/Noise Modeling Combined with Dynamic Variance Adaptation, (2011), pp. 12–17. http://spandh.dcs.shef.ac.uk/projects/chime/workshop/papers/pS21_delcroix.pdf
60. F Xiong, N Moritz, R Rehr, J Anemuller, BT Meyer, T Gerkmann, S Doclo, S Goetze, in *Proc. REVERB Workshop*. Robust ASR in Reverberant Environments Using Temporal Cepstrum Smoothing for Speech Enhancement and an Amplitude Modulation Filterbank for Feature Extraction, (2014). http://reverb2014.dereverberation.com/workshop/reverb2014-papers/1569899061.pdf
61. H-G Hirsch, HMM adaptation for applications in telecommunication. Speech Commun. **34**(1-2), 127–139 (2001)
62. H Jiang, K Hirose, Q Huo, Robust speech recognition based on a Bayesian prediction approach. IEEE Trans. Speech Audio Process. **7**(4), 426–440 (1999)
63. J-T Chien, Linear regression based Bayesian predictive classification for speech recognition. IEEE Trans. Speech Audio Process. **11**(1), 70–79 (2003)
64. A Nadas, D Nahamoo, MA Picheny, Speech recognition using noise-adaptive prototypes. IEEE Trans. Audio, Speech, Lang. Process. **37**(10), 1495–1503 (1989)
65. R Maas, A Sehr, M Gugat, W Kellermann, in *Proc. European Signal Processing Conf. (EUSIPCO)*. A Highly Efficient Optimization Scheme for REMOS-Based Distant-Talking Speech Recognition (IEEE, New Jersey, USA, 2010), pp. 1983–1987
66. SC Schwartz, YS Yeh, On the distribution function and moments of power sums with log-normal components. Bell Syst. Tech. Journal. **61**(7), 1441–1462 (1982)
67. NC Beaulieu, AA Abu-Dayya, PJ McLane, Estimating the distribution of a sum of independent lognormal random variables. IEEE Trans. Commun. **43**(12), 2869–2873 (1995)
68. CK Raut, T Nishimoto, S Sagayam, in *Proc. Interspeech*. Model Composition by Lagrange Polynomial Approximation for Robust Speech Recognition in Noisy Environment (ISCA, Baixas, France, 2004)
69. NC Beaulieu, Q Xie, An optimal lognormal approximation to lognormal sum distributions. IEEE Trans. Veh. Technol. **53**(2), 479–489 (2004)
70. JR Hershey, SJ Rennie, JL Roux, in *Techniques for Noise Robustness in Automatic Speech Recognition*, ed. by T Virtanen, R Singh, and B Raj. Factorial Models for Noise Robust Speech Recognition (Wiley, UK, 2013), pp. 311–345
71. D Yu, L Deng, *Automatic Speech Recognition—A Deep Learning Approach*. (Springer, London, 2015)
72. ML Seltzer, D Yu, Y Wang, in *Proc. ICASSP*. An Investigation of Deep Neural Networks for Noise Robust Speech Recognition (IEEE, New Jersey, USA, 2013), pp. 7398–7402
73. L Deng, J Li, J-T Huang, K Yao, D Yu, F Seide, M Seltzer, G Zweig, X He, J Williams, Y Gong, A Acero, in *Proc. ICASSP*. Recent Advances in Deep Learning for Speech Research at Microsoft (IEEE, New Jersey, USA, 2013), pp. 8604–8608
74. R Gemello, F Mana, S Scanzio, P Laface, R De Mori, in *Proc. ICASSP*. Adaptation of Hybrid ANN/HMM Models Using Linear Hidden Transformations and Conservative Training (IEEE, New Jersey, USA, 2006), pp. 1189–1192
75. F Seide, G Li, X Chen, D Yu, in *Proc. ASRU*. Feature Engineering in Context-Dependent Deep Neural Networks for Conversational Speech Transcription (IEEE, New Jersey, USA, 2011), pp. 24–29
76. K Yao, D Yu, F Seide, H Su, L Deng, Y Gong, in *Proc. SLT*. Adaptation of Context-Dependent Deep Neural Networks for Automatic Speech Recognition (IEEE, New Jersey, USA, 2012), pp. 366–369
77. G Saon, H Soltau, D Nahamoo, M Picheny, in *Proc. ASRU*. Speaker Adaptation of Neural Network Acoustic Models Using I-Vectors (IEEE, New Jersey, USA, 2013), pp. 55–59
78. RF Astudillo, JP da Silva Neto, in *Proc. Interspeech*. Propagation of Uncertainty Through Multilayer Perceptrons for Robust Automatic Speech Recognition (ISCA, Baixas, France, 2011), pp. 461–464