# A Hierarchical Bayesian Approach to Modeling Heterogeneity in Speech Quality Assessment

Iman Mossavat, *Student Member, IEEE*, Petko N. Petkov, *Student Member, IEEE*, W. Bastiaan Kleijn, *Fellow, IEEE*, and Oliver Amft, *Member, IEEE*

***Abstract*—The development of objective speech quality measures generally involves fitting a model to subjective rating data. A typical data set comprises ratings generated by listening tests performed in different languages and across different laboratories. These factors as well as others, such as the sex and age of the talker, influence the subjective ratings and result in data heterogeneity. We use a linear hierarchical Bayes (HB) structure to account for heterogeneity. To make the structure effective, we develop a variational Bayesian inference for the linear HB structure that approximates not only the posterior over the model parameters, but also the model evidence. Using the approximate model evidence we are able to study and exploit the heterogeneity inducing factors in the Bayesian framework. The new approach yields a simple linear predictor with state-of-the-art predictive performance. Our experiments show that the new method compares favorably with systems based on more complex predictor structures such as ITU-T recommendation P.563, Bayesian MARS, and Gaussian processes.**

***Index Terms*—Heterogeneity, hierarchical Bayesian, multi-task learning, non-intrusive, quality of service, single-ended, speech quality, variational inference.**

## I. Introduction

MAINTAINING the quality of service in telecommunication networks is an important task that requires reliable estimation of speech quality. Subjective listening tests are the most accurate way to assess the speech quality. Unfortunately, this accuracy comes at the price of human labor as well as stringent requirements on testing conditions [1]–[4]. A more fundamental limitation of listening tests is that subjective speech testing cannot be used in real-time quality assessment (QA) applications such as monitoring the quality of live calls. Objective speech QA measures are algorithms that aim to estimate the average of subjective quality ratings and were originally developed to replace listening tests. Improving the prediction accuracy of objective QA measures remains a major research focus since the advent of first quality models in 1980s [1]. Speech QA has become an increasingly complex task today as communication networks have become highly heterogeneous and transmission medium, network protocols, and coding algorithms all may have an impact on the final quality.

QA measures can be divided into two categories based on their inputs. *Intrusive* (double-ended) measures require a reference speech signal in addition to the distorted speech, where the reference is a clean version of the distorted speech. The current "state-of-the-art" in intrusive measures is represented by the ITU-T P.862 standard, also referred to as perceptual evaluation of speech quality (PESQ) [5], [6]. In applications such as monitoring the quality of live-calls, the reference signal used by intrusive measures is not available or costly to provide. Thus, the need for a reference signal limits the applicability of intrusive measures. *Non-intrusive* (single-ended) measures, on the other hand, assess the quality based on the distorted signal only. The current "state-of-the-art" in non-intrusive QA is represented by ITU-T P.563 standard [7]. The design of non-intrusive QA measures is more involved than intrusive QA measures because the reference signal is not available.

To develop algorithms that estimate speech quality as "perceived" by human subjects, quality rating data sets provided by listening tests are used. Data sets typically comprise speech signals from different languages, which are rated by native speakers in different laboratories across the world. The speech signals are distorted by conditions similar to those occurring in telecommunication systems. Different distortion conditions affect the speech quality differently. The variation of attributes such as the language, the testing laboratory, or the distortion conditions may considerably affect the statistical properties of the data such as the distributions of speech features, and the joint distribution of speech features and the quality ratings, creating data *heterogeneity*. It is a common practice to ignore the variations due to heterogeneity inducing attributes, and *pool* the data coming from different sources into one training set. However, modeling heterogeneous data by pooling is potentially detrimental to prediction performance [8], [9]. Thus, it is important to investigate the advantages of taking heterogeneity into account when modeling quality rating data.

In this work, we concentrate on the problem of data heterogeneity in the context of non-intrusive QA. Our study focuses on the fitting of the quality models, i.e., learning the mapping of features to quality estimates from the training data. The nov-

Manuscript received November 09, 2010; revised March 14, 2011; accepted May 10, 2011. Date of publication June 02, 2011; date of current version November 09, 2011. The work of I. Mossavat was supported by STW project 07605: Personalization of Hearing Aids Through Bayesian Preference Elicitation (HearClip). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Engin Erzin.

I. Mossavat and O. Amft are with the Signal Processing Systems Group, Department of Electrical Engineering, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands. (e-mail: i.mossavat@tue.nl; o.amft@tue.nl; amft@tue.nl).

P. N. Petkov is with the Sound and Image Processing Lab, Department of Electrical Engineering, KTH–Royal Institute of Technology, SE-100 44 Stockholm, Sweden. (e-mail: petko.petkov@ee.kth.se).

W. B. Kleijn is with the Sound and Image Processing Lab, Department of Electrical Engineering, KTH–Royal Institute of Technology, SE-100 44 Stockholm, Sweden, and also with the School of Engineering and Computer Science, Victoria University of Wellington, Wellington 6012, New Zealand (e-mail: bastiaan.kleijn@ee.kth.se; bastiaan.kleijn@ecs.vuw.ac.nz).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2158421

elty of our approach lies in considering the data heterogeneity explicitly via multi-task learning [8]. Instead of pooling different types of data into one training set, multi-task learning uses data attributes such as language to create subsets of data called *tasks*. Tasks are modeled jointly along with an interaction model that learns the similarities of task models. The Bayesian framework offers a natural way to create a multi task learning algorithm, called the hierarchical Bayes (HB) structure [10]. We develop a linear HB structure and present an efficient fully Bayesian approximate inference algorithm for parameter estimation. In addition to estimating the model parameter, the full Bayesian methodology allows comparison of different task configurations, i.e., how data attributes should be used to divide the training set into tasks. We present our algorithm in detail, such that it is possible to implement and perform experiments on private data sets. Based on experimental results, our approach results in performance improvement as well as substantial simplification of the predictor complexity.

To analyze the predictive performance of linear HB structure-based QA measures, we use the ITU-T Supplement 23, the ITU coded-speech data set [11]. Supplement 23 is chosen because it is heterogeneous, publicly available, and widely adopted. Our approach to handling heterogeneity in speech data sets is not limited to Supplement 23, and can be applied to designing feature mappings for other heterogeneous speech quality rating data sets.

We compare the performance of our linear HB structure based QA measure against the state-of-the-art P.563. We also compare our algorithm with two regression algorithms that are trained on the pooled Supplement 23 data: Bayesian multivariate adaptive regression splines (BMARS) [12] and Gaussian process regression (GPR) [13]. The linear HB structure yields a *linear* QA feature mapping, which provides competitive performance with a predictor complexity that is substantially smaller than P.563, BMARS, and GPR algorithms.

This paper is organized as follows: related work is briefly reviewed in Section II, and the basic concepts of multi-task learning are reviewed in Section III. In Section IV, we first review the generic HB structure, where the full Bayesian prediction, inference and model comparison for the HB structure is explained. We continue Section IV by presenting our linear HB structure. In Section V, the evaluation procedure is explained. The results are given in Section VI. Discussion on the results is given in Section VII, and our conclusion are summarized in Section VIII. In Appendix A, we present the details of our linear HB structure inference algorithm. In Appendix B, we give the details of BMARS and GPR algorithms settings used in our experimentations.

## II. Related Work

Speech quality estimation attracted substantial research effort. A few examples of intrusive measures are the early work of Karjalainen [1], weighted-slope spectral distance (WSSD) [14], bark spectral distance (BSD) [15], [16], perceptual speech quality measure (PSQM) [17], and measuring normalizing blocks (MNB) [18], [19].

Some examples of non-intrusive measures are the early work of Liang and Kubichek [20], the QA measure based on vocal-tract models [21], ANIQUE [22] and ANIQUE+ [23], algorithms by Falk *et al.* based on machine learning concepts [24]–[26], and the work of Chen and Parsa based on Bayesian inference [27]. The ITU-T P.563 standard [7] is based on three algorithms: non-intrusive speech quality assessment (NiQA) [28], [29]; non-intrusive network assessment (NiNA) [30]; and perceptual single-sided speech quality measure (P3SQM).

The feature mapping design for non-intrusive speech quality estimation becomes more challenging in the face of heterogeneity as the reference signal is absent. The most prominent example is the ITU-T standard P.563 that uses a divide-and-conquer (DaC) strategy in its feature mapping to handle heterogeneity in distortions occurring in telecommunication systems. The DaC is realized via an algorithm called the dominant distortion classification and and perceptual weighting (DDCPW), which uses speech features to determine the type of speech distortion and the sex of the talker. As a result, handling heterogeneity results in a complex mapping which is fitted to the pooled data. Attributes of the training data such as the test laboratory, or language are not easy to express using speech features, and are often ignored by the DaC. The DaC strategy is limited to heterogeneity caused by factors that can be expressed using speech features such as different telecommunication network distortions, or sex of the talker.

Machine learning and statistical methods have been used to develop complex mappings: multivariate adaptive regression splines (MARS) [31] was used in [32]–[34]; self-organizing maps were used in [35]; Gaussian mixture models were used in [24]; and, support vector regression were used in [36]; in [37] we used a DaC strategy based on the Bayesian mixture of experts model of [38]. To fit the model, all aforementioned methods pool the heterogeneous data into one set, ignoring the undesired variations in the data.

## III. Learning From Multiple Tasks

Multi-task learning, that is, learning multiple related predictive functions jointly, aims to share knowledge gained from related scenarios [8], [9]. The key aspect of multi-task learning is that it deals with heterogeneity during the training and yields simple predictors during deployment, which operate globally, i.e., on all scenarios.

In the training phase, we divide the data according to attributes that cause the variations in the data distribution into "tasks" (data subsets), for example in our data set the laboratory of origin or distortion condition may induce heterogeneity in the data. It is important to distinguish speech features from data attributes in our work. Attributes are labels on the training data, such as "French," "male speaker," or "CNET laboratory." Speech features form a compact representation of the signal. For quality estimation, a good feature set should preserve all the information that is vital for prediction speech quality.

We study the significance of heterogeneity caused by attributes both from the performance perspective, as well as from the Bayesian viewpoint. We train separate task-specific models simultaneously by "linking" these models in our training algorithm such that they regularize each other and gain statistical strength. In other words, we jointly learn different but "related" models (quality measures) for different

tasks, such that task-specific models can interact at a higher level. To perform multi-task learning, we use the *linear* HB structure. In our linear HB structure, task-specific models sit at the bottom of the hierarchy. All task-specific models are linear models, but with different distributions on the feature weights. At the second layer, a prior distribution over parameters of the task-specific models is placed as a task-interaction model. The task-interaction model is meant to capture the similarities between parameters of the task-specific models. To compare different ways of forming tasks using the Bayesian theory, our algorithm approximates the log-model-evidence, which is the logarithm of the probability of the data given the model. Task configurations are compared against each other as alternative models.

It is important to note the difference between the divide-and-conquer methodology exploited in P.563 and our earlier work [37] with the new multi-task approach: in P.563 some features are used to classify the signals during the algorithm *deployment* stage, whereas our multi-task approach uses the attributes only during the training stage to split the data. That is, our new approach does not divide the features space to predict the speech quality using local predictors as in P.563 and [37]. The generic HB structure is described in Section IV. We specify the linear HB structure in Section IV-D, which is used in our experiments.

## IV. Hierarchical Bayes Structure

Linear HB structures are applied widely in different applications [10], [39]. Our linear model is most similar to the one used in [39], but we develop a variational Bayesian inference for training our model, whereas in [39] the maximum-likelihood principle is used for training. Our inference method allows to approximate the log-mode-evidence, which in turn, allows us to study the heterogeneity caused by various attributes in the Bayesian framework.

### A. Generic HB Structure

Consider a data set $\mathcal{D}$, which is divided into $m$ subsets (tasks) $\mathcal{D}_l$ for $l = 1,\ldots,m$. From this point forward we use $l$ and $i$ to refer to the task and data point indices, respectively. The $l$th task is denoted by $\mathcal{D}_l = \{(\boldsymbol{\psi}_{il}, y_{il}) | i = 1,\ldots,n_l\}$, where the number of data points, the $i$th speech signal, and the corresponding MOS is denoted by $n_l$, $\boldsymbol{\psi}_{il}$, and $y_{il}$, respectively. The division of the data set into tasks is specified according to the model denoted by $\mathcal{M}$.

Using the aforementioned notation, we use the following parametric probabilistic model, called the task specific model, for the data points in the $l$th task:

$$\Pr(y_{il}|\boldsymbol{\psi}_{il}) = \mathrm{F}(y_{il}|\boldsymbol{\psi}_{il}, \boldsymbol{\omega}_l) \tag{1}$$

where we used the vector $\boldsymbol{\omega}_l$ to denote the task parameter vector. Task specific models are placed at the bottom of the HB structure. At the second layer of the HB structure, the task interaction model is specified, which is a probabilistic model over the space of task parameter vector, i.e.,

$$\Pr(\boldsymbol{\omega}_l) = \mathrm{G}(\boldsymbol{\omega}_l|\boldsymbol{\phi}). \tag{2}$$

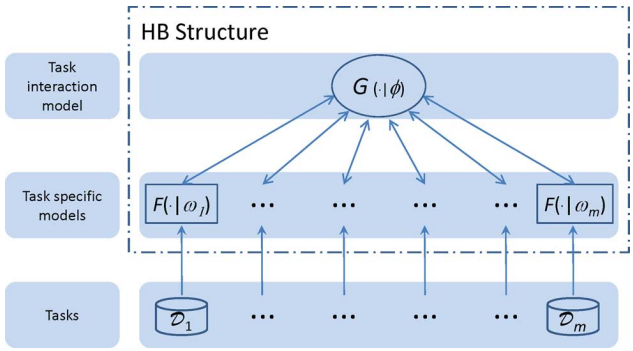



Fig. 1. Multi-task learning using the hierarchical Bayes structure: data is divided into $m$ tasks $\mathcal{D}_l$ for $l = 1,\ldots,m$ using the attributes specified by the task configuration. Task-specific models are described by parametric distributions $\mathrm{F}(.|\boldsymbol{\omega}_l)$ with different task parameter vectors $\boldsymbol{\omega}_l$. Task parameters are linked using a parametric task-interaction model $G(\cdot|\boldsymbol{\phi})$ at the top of the hierarchy, where G is the prior distribution over task parameter vector. Task interaction parameter vector $\boldsymbol{\phi}$ is learned simultaneously with vectors $\boldsymbol{\omega}_l$ as illustrated by two-sided arrows, resulting in the desired task interaction.

The task interaction model $\mathrm{G}(\boldsymbol{\omega}_l|\boldsymbol{\phi})$ is a "prior" distribution over task parameter vector $\boldsymbol{\omega}_l$ conditioned on the interaction parameter denoted by the vector $\boldsymbol{\phi}$. Task interaction model $\mathrm{G}(\cdot|\boldsymbol{\phi})$ represents the similarities shared by the parameters of task specific models, and learning $\boldsymbol{\phi}$ amounts to learning the "prior" over $\boldsymbol{\omega}_l$. Learning $\boldsymbol{\phi}$ *jointly* with $\boldsymbol{\omega}_l$ results in the desired interaction between task specific models, which is our primary motivation in using the HB structure. Fig. 1 illustrates the architecture of the HB structure.

### B. Prediction

Full Bayesian prediction amounts to computing the predictive distribution $\Pr(y|\boldsymbol{\psi}, \mathcal{D}, \mathcal{M})$, where the signal quality $y$ is conditioned only on the feature vector $\boldsymbol{\psi}$ and the attributes are not present in the conditioning, because we are interested in using the attributes only in the training phase. Other terms in the conditioning include previous observations, i.e., our data set $\mathcal{D}$ and the model $\mathcal{M}$. The tasks are defined by the model $\mathcal{M}$. The predictive distribution for the HB structure is

$$\Pr(y|\boldsymbol{\psi}, \mathcal{D}, \mathcal{M}) = \int \mathrm{F}(y|\boldsymbol{\psi}, \boldsymbol{\omega})\ \mathrm{G}(\boldsymbol{\omega}|\boldsymbol{\phi}) \Pr(\boldsymbol{\phi}|\mathcal{D}, \mathcal{M}) \mathrm{d}\boldsymbol{\omega}\ \mathrm{d}\boldsymbol{\phi} \tag{3}$$

where distributions F and G are defined in (1) and (2), respectively. The posterior $\Pr(\boldsymbol{\phi}|\mathcal{D}, \mathcal{M})$ is computed via the Bayesian inference. over the interaction parameter vector $\boldsymbol{\phi}$.

Based on (3), it is possible to explain how HB structure is used to learn our speech quality estimator. To compute the predictive distribution, we average the task specific model $\mathrm{F}(y|\boldsymbol{\psi}, \boldsymbol{\omega})$, where the unknown task parameter vector $\boldsymbol{\omega}$ is sampled from the distribution $\mathrm{G}(\boldsymbol{\omega}|\boldsymbol{\phi})$, whose unknown interaction parameter $\boldsymbol{\phi}$ is in turn sampled from the interaction parameter posterior $\Pr(\boldsymbol{\phi}|\mathcal{D}, \mathcal{M})$. The interaction parameter posterior is learned from the data as explained shortly. Note that the task specific parameters $\boldsymbol{\omega}_l$ used in the learning stage are discarded from the posterior by marginalization.

We use the mean of the predictive distribution as our point quality estimate. Computing (3) is complex in general. For the linear HB structure in the next section, we show how to efficiently compute the mean by a weighted sum of features.

TABLE I
TASK CONFIGURATIONS BASED ON DIFFERENT DATA ATTRIBUTES. THE ATTRIBUTES ARE: SUPPLEMENT 23 EXPERIMENT TYPE (TWO OPTIONS), SUPPLEMENT 23 DATABASE (SEVEN), AND SEX OF THE TALKER (TWO)

| Configuration | Considered attributes | Number of tasks ($m$) |
|---|---|---|
| SD | Sex of the talker-Database | 14 |
| SE | Sex of the talker-Experiment | 4 |
| S | Sex of the talker | 2 |
| D | Database | 7 |
| E | Experiment | 2 |

### C. Inference and Model Evidence

Based on (3), the interaction parameter posterior $\Pr(\boldsymbol{\phi}|\mathcal{D},\mathcal{M})$ is required for prediction; thus, we write the Bayes rule for the HB structure as follows:

$$\Pr(\boldsymbol{\phi}|\mathcal{D},\mathcal{M}) = \frac{\boldsymbol{\pi}(\boldsymbol{\phi}|\mathcal{M})\Pr(\mathcal{D}|\boldsymbol{\phi},\mathcal{M})}{\Pr(\mathcal{D}|\mathcal{M})} \tag{4}$$

where $\boldsymbol{\pi}(\boldsymbol{\phi}|\mathcal{M})$ denoted the prior on $\boldsymbol{\phi}$, which is specified by the model $\mathcal{M}$. The *likelihood* of the HB structure is written as

$$\Pr(\mathcal{D}|\boldsymbol{\phi},\mathcal{M}) = \prod_{l=1}^{m} \Pr(\mathcal{D}_l|\boldsymbol{\phi}) \tag{5}$$

where the tasks are defined according to the model $\mathcal{M}$, and

$$\Pr(\mathcal{D}_l|\boldsymbol{\phi}) = \int \left( \prod_{i=1}^{n_l} F(y_{il}|\boldsymbol{\psi}_{il},\boldsymbol{\omega}_l) \right) G(\boldsymbol{\omega}_l|\boldsymbol{\phi})\, \mathrm{d}\boldsymbol{\omega}_l. \tag{6}$$

The *model evidence* for the generic HB structure is computed as

$$\Pr(\mathcal{D}|\mathcal{M}) = \int \Pr(\mathcal{D}|\boldsymbol{\phi},\mathcal{M})\boldsymbol{\pi}(\boldsymbol{\phi}|\mathcal{M})\, \mathrm{d}\boldsymbol{\phi} \tag{7}$$

where the likelihood of the generic HB structure is defined in (5) and $\boldsymbol{\pi}$ denotes the prior on $\boldsymbol{\phi}$. The model evidence serves as the normalization constant in (4) and is not necessary for prediction; however, it plays the central role in selection of the model $\mathcal{M}$, which includes the specification of the tasks as well as the prior $\boldsymbol{\pi}$ on the interaction parameter. We assume all models specify the same prior on the interaction parameter, so Bayesian model selection amount to Bayesian selection of the task configuration. Different attributes such as language, sex of the talker, laboratory, and distortion condition are possible sources of heterogeneity. In our multi-task learning approach, data attributes are used to specify the task design. Using the Bayesian framework, we use the model-evidence as the tool for comparing different combinations of attributes in the design of tasks. The task configurations in our experiments are given in Table I in the next section.

Full Bayesian inference and model selection are complex in general. In the following, we present a linear HB structure, as well as efficient approximation methods for inference, model selection and prediction.

### D. Linear HB Structure

The Bayesian *linear* HB structure is presented by specifying the task specific model F, task-interaction model G, and the prior over interaction parameters $\boldsymbol{\pi}$. The models and the prior $\boldsymbol{\pi}$ are chosen such that efficient approximate Bayesian inference based on the mean-field variational methods is possible [40].

Linear task specific models are defines as

$$\Pr(y_{il}|\boldsymbol{\psi}_{il}) = \mathcal{N}(y_{il}|\boldsymbol{\omega}_l^{\mathrm{T}}\boldsymbol{\psi}_{il}, \lambda^{-1}). \tag{8}$$

where $i$ and $l$ denote the signal and the task index, and a zero-mean Gaussian noise with precision (inverse of variance) $\lambda$ is assumed. The task interaction model is set to

$$\Pr(\boldsymbol{\omega}_l) = \mathcal{N}(\boldsymbol{\omega}_l|W, (\lambda\Lambda)^{-1}). \tag{9}$$

where $W$ is a $D$-dimensional vector and $\Lambda$ is a $D \times D$ positive-definite matrix, with $D$ being the dimensionality of feature vectors, and a zero-mean Gaussian noise with precision matrix $\lambda\Lambda$ is assumed. The presence of $\lambda$ in the precision matrix of (9) is required for developing our variational inference algorithm. Thus, the interaction parameters in the linear HB structure comprise $\{W, \Lambda\}$. The prior over interaction parameters is

$$\begin{aligned} \Pr(W|\Lambda) &= \mathcal{N}(W|W_0, (\alpha\Lambda)^{-1}) \\ \Pr(\Lambda) &= \mathcal{W}(\Lambda|\tau_0, \Sigma_0) \end{aligned} \tag{10}$$

where $W_0 = 0$ if no prior knowledge exists, and $\mathcal{W}$ is the Wishart distribution with hyper-parameters $\tau_0$ and $\Sigma_0$ denoting the degrees of freedom and the scale matrix, respectively. We set the hyper-parameters to values corresponding to a weak (non-informative) prior, i.e., $\tau_0 = D$ and $\Sigma_0 = \boldsymbol{I}(D)$, where $\boldsymbol{I}(D)$ is the identity matrix of dimensionality $D$. The choice of identity matrix implies assuming no correlation in the (10). Respecting the conjugate structure in the model, the prior over $\lambda$ and $\alpha$ is chosen to conform to a Gamma distribution as follows:

$$\begin{aligned} \Pr(\lambda|\mathcal{M}) &= \mathrm{Gamma}(\lambda|a_0^{\lambda}, b_0^{\lambda}) \\ \Pr(\alpha|\mathcal{M}) &= \mathrm{Gamma}(\alpha|a_0^{\alpha}, b_0^{\alpha}). \end{aligned} \tag{11}$$

where the hyper-priors are set in the model $\mathcal{M}$ which correspond to wide non-informative priors as following: the model sets $a_0^{\lambda,\alpha} = 10^{-2}$ and $b_0^{\lambda,\alpha} = 10^{-4}$.

Appendix A gives the details of the approximate inference for our linear model, where the posterior over variation parameters are computed. Based on the results in Appendix A, the mean of the predictive distribution in (3) is computed to be

$$\mathbb{E}[y|\boldsymbol{\psi}] = \boldsymbol{\mu}_W^{\mathrm{T}}\boldsymbol{\psi} \tag{12}$$

where $\boldsymbol{\mu}_W$ is the mean of the *posterior* on $W$, which is the average feature weight in task specific models.

In contrast to P.563 [41] as well as our earlier work based on the Bayesian mixture of experts [37], where feature mapping comprised local predictors, the HB predictor of (12) is global, i.e., a single linear mapping of the features works as a predictor over the entire space of input vectors. Thus, we do not need to divide the feature space into subspaces.

### E. Approximate Inference

A variational approximation to the posterior based on the *mean-field* theory [40] is used in this work. Let $\boldsymbol{\theta}$ denote the set of linear HB structure parameters:

$\boldsymbol{\theta} = \{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_m, W, \Lambda, \lambda, \alpha\}$. For any arbitrary distribution $q(\boldsymbol{\theta})$, it is possible to write the log-model-evidence as follows:

$$\ln \Pr(\mathcal{D}|\mathcal{M}) = \mathcal{L}(q) + \mathrm{KL}(q\|p) \tag{13}$$

where

$$\mathcal{L}(q) = \int q(\boldsymbol{\theta}) \ln \frac{\Pr(\boldsymbol{\theta}, \mathcal{D}|\mathcal{M})}{\Pr(\boldsymbol{\theta}|\mathcal{M})} \mathrm{d}\boldsymbol{\theta} \tag{14}$$

and $\mathrm{KL}(q\|p)$ is the Kullback–Leibler (KL) divergence between distributions $q(\boldsymbol{\theta})$ and the posterior $\Pr(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})$. The divergence is always non-negative; thus, $\mathcal{L}(q)$ is a variational *lower-bound* to log-model-evidence. In other words

$$\ln \Pr(\mathcal{D}|\mathcal{M}) \geq \mathcal{L}(q) \tag{15}$$

where the equality occurs if and only if the variational distribution $q(\boldsymbol{\theta})$ matches the posterior $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})$.

Approximating the posterior by $q(\boldsymbol{\theta})$ amounts to reducing the KL divergence $\mathrm{KL}(q\|p)$, which is equivalent to maximizing the variational lower-bound $\mathcal{L}(q)$ according to (13). An efficient iterative algorithm for selecting the optimal $q(\boldsymbol{\theta})$ from a certain family of functions exists. More details are give in Appendix A.

## V. Evaluation Procedure

To demonstrate the effectiveness of considering heterogeneity, we compare our linear HB structure performance against two data-driven regression methods that are trained on pooled data: BMARS [12] and GPR [13]. Both algorithms are data-driven (non-parametric) and have been successfully used in various data modeling and prediction problems. The details of the BMARS and GPR settings are given in Appendix B.

For the linear HB structure the model $\mathcal{M}$ in (3) specifies the task configuration and prior hyper-parameters, which are explained shortly in this section.

### A. Supplement 23: The ITU-T Coded-Speech Data Set

ITU-T coded-speech data set, Supplement 23, comprises three experiments [11], [42]. Experiments 1 and 3 use an absolute category rating (ACR) protocol and experiment 2 uses a comparative category rating (CCR) protocol. We choose the ACR test results of experiments 1 and 3 for our performance analysis as this has been often used to evaluate the performance of non-intrusive QA algorithms, e.g., in P.563 [41]. In the ACR method, human listeners (subjects) are required to grade the speech files on a discrete opinion scale: "Excellent," "Good," "Fair," "Poor," "Bad." For each speech file in Supplement 23, votes are averaged over 24 subjects and referred to as the mean opinion score (MOS) [4].

Experiments 1 and 3 of Supplement 23 are divided into seven "databases" coming from four different laboratories/languages: BNR (English), CNET (French), CSELT (Italian), and NTT (Japanese). Experiment 1 includes 44 speech coding distortion conditions, while experiment 3 includes 50 channel impairment conditions, such as various conditions of introducing noise to coded-speech. Our data set comprises 1328 narrow-band speech files. Supplement 23 is a heterogeneous data set as it covers four languages and several distortion conditions. The details of the distortion conditions are given in [42].

### B. Feature Set

We use the feature set of the ITU-T P.563, which comprises 43 narrowband features extracted based on the following three principles [41]: the first group of features are extracted by using a vocal tract model and LPC analysis. The second group of features is extracted by reconstruction of a quasi-clean reference signal and using a modified PESQ algorithm [5], [6], and finally the distortion-specific features are extracted based on models of channel distortions. These features include measurements of the noise, and detection of temporal clipping and robotization.

While our feature set and data set are dealing with narrowband speech, our method to learn feature mappings from heterogeneous data is conceptually not limited to narrowband speech. It is straightforward to implement and test our method for other data sets and feature sets based on the implementation details provided in Appendix A.

### C. Task Definitions for the HB Structure

The first step in modeling heterogeneity using the HB structure is dividing the data set into smaller sets called tasks. Each task configuration corresponds to a hypothesis on what attributes are causing heterogeneity. For example, as explained earlier in Section V-A, the seven databases in Supplement 23 are coming from different laboratories (different languages) and different experiments. Another possible heterogeneity inducing attribute is the sex of the talker, as assumed in the P.563 standard.

The attributes and task configurations used in this work are given in Table I. Each configuration corresponds to a separate model $\mathcal{M}$ in (4). Our method allows a Bayesian model selection by efficiently approximating the log-model-evidence with $\mathcal{L}(q)$ defined in (14).

Tasks are used only for training, and assumed to be unknown to the predictor. For example, sex of the talker is considered as a heterogeneity inducing attribute, which is only used during the training. Thus, the same predictor is used for male and female speakers. This is contrast with P.563, where the algorithm detects the sex of the talker and chooses the corresponding *local* predictor.

### D. Prior Hyper-Parameters

Linear HB structure priors are specified by hyper-parameters in (10) and (11). We set the hyper-parameters by performing the Bayesian inference twice (See Chapter 5 of Gelman *et al.* [10]). For the first inference, we set the hyper-parameters such that broad, non-informative priors are obtained, and approximate the posterior. Then we use the approximate posterior to fine-tune the priors and perform the final inference. In more detail, we set $W_0 = \sum_l \mathbb{E}[\boldsymbol{\omega}_l]/m$, and we set $a_0^\lambda$, $b_0^\lambda$, $a_0^\alpha$, and $b_0^\alpha$ to match the priors over $\lambda$ and $\alpha$ with their posteriors. We keep $\tau_0$, and $\Sigma_0$ intact. Our experiments indicates that the two-round inference scheme consistently improves the prediction accuracy over the case where non-informative priors are used.

TABLE II
PCC AND RMSE ON THE TEST DATA, CALCULATED PER-SPEECH-FILE: (a), (b), AND PER-CONDITION (c), (d). EACH DATABASE IN SUPPLEMENT 23 CORRESPONDS TO A LAB AND AN EXPERIMENT, WHERE X1 AND X3 DENOTE EXPERIMENTS 1 AND 3, RESPECTIVELY. TASK CONFIGURATION SE WAS USED FOR THIS ANALYSIS. (a) PCC (PER-SPEECH-FILE). (b) RMSE (PER-SPEECH-FILE). (c) PCC (PER-CONDITION). (d) RMSE (PER-CONDITION)

| Database | HB | GPR | BMARS | P.563 |
|---|---|---|---|---|
| BNR-X1 | 0.79 | 0.82 | 0.84 | 0.79 |
| BNR-X3 | 0.73 | 0.79 | 0.82 | 0.78 |
| CNET-X1 | 0.76 | 0.81 | 0.78 | 0.76 |
| CNET-X3 | 0.85 | 0.71 | 0.71 | 0.77 |
| NTT-X1 | 0.79 | 0.82 | 0.81 | 0.65 |
| NTT-X3 | 0.85 | 0.79 | 0.79 | 0.82 |
| CSELT-X3 | 0.83 | 0.73 | 0.72 | 0.76 |
| Mean | 0.80 | 0.78 | 0.78 | 0.76 |

(a)

| Database | HB | GPR | BMARS | P.563 |
|---|---|---|---|---|
| BNR-X1 | 0.69 | 0.53 | 0.55 | 0.53 |
| BNR-X3 | 0.59 | 0.50 | 0.50 | 0.50 |
| CNET-X1 | 0.52 | 0.45 | 0.50 | 0.51 |
| CNET-X3 | 0.52 | 0.74 | 0.77 | 0.58 |
| NTT-X1 | 0.51 | 0.46 | 0.46 | 0.61 |
| NTT-X3 | 0.51 | 0.69 | 0.69 | 0.61 |
| CSELT-X3 | 0.43 | 0.81 | 0.92 | 0.70 |
| Mean | 0.55 | 0.60 | 0.63 | 0.58 |

(b)

| Database | HB | GPR | BMARS | P.563 |
|---|---|---|---|---|
| BNR-X1 | 0.87 | 0.92 | 0.94 | 0.92 |
| BNR-X3 | 0.86 | 0.91 | 0.92 | 0.91 |
| CNET-X1 | 0.90 | 0.90 | 0.87 | 0.89 |
| CNET-X3 | 0.93 | 0.79 | 0.81 | 0.87 |
| NTT-X1 | 0.95 | 0.93 | 0.95 | 0.80 |
| NTT-X3 | 0.96 | 0.91 | 0.89 | 0.93 |
| CSELT-X3 | 0.93 | 0.83 | 0.83 | 0.84 |
| Mean | 0.91 | 0.88 | 0.89 | 0.88 |

(c)

| Database | HB | GPR | BMARS | P.563 |
|---|---|---|---|---|
| BNR-X1 | 0.41 | 0.30 | 0.27 | 0.31 |
| BNR-X3 | 0.35 | 0.29 | 0.28 | 0.30 |
| CNET-X1 | 0.32 | 0.33 | 0.38 | 0.34 |
| CNET-X3 | 0.25 | 0.42 | 0.40 | 0.34 |
| NTT-X1 | 0.20 | 0.23 | 0.20 | 0.38 |
| NTT-X3 | 0.21 | 0.28 | 0.31 | 0.25 |
| CSELT-X3 | 0.26 | 0.47 | 0.46 | 0.44 |
| Mean | 0.29 | 0.33 | 0.33 | 0.34 |

(d)

### *E. Evaluation Metric*

We use cross-validation and report Pearson correlation-coefficient (PCC) and root-mean-squared error (RMSE). PCC and RMSE are the established measures for evaluating the performance of QA measures. We report both per-file results, as well as per-condition results. The per-condition metrics are computed by first averaging the model predictions as well as the test set MOS over conditions, and then computing the PCC and RMSE on the averaged results [37]. We apply a *monotonic* third-order polynomial as recommended in [41] to map the condition-averaged predictions to condition-averaged MOS. For the linear HB the per-condition results are computed with and without the polynomial mapping, and it is observed that the polynomial adds small improvement to the condition-averaged performance.

## VI. RESULTS

### *A. Prediction Performance Results*

The following cross-validation procedure is used to compare the prediction accuracy of learning methods: six databases of Supplement 23 are used as the training set, and the remaining database is used as the test set. The PCC and RMSE are computed on the test set. MOS is the prediction target variable. For BMARS and GPR the six training databases are pooled together to form the training set, while for the linear HB structure tasks are formed according to the definitions in the previous section.

Table II reports the per-file and per-condition PCC and RMSE results for the following methods: linear HB (with task configuration SE), BMARS, GPR, and the P.563. As results indicate, the linear predictor trained using the HB structure (with SE tasks) performs slightly better than the other methods by achieving higher per-file PCC and lower RMSE. Note the reduced spread of PCC and RMSE values for linear HB in comparison with other methods. For linear HB the PCC is always larger than 0.73, and the RMSE is always below 0.69. For GPR and BMARS PCC reaches 0.71. The RMSE shows a larger performance gap: the maximum RMSE for GPR and BMARS are 0.81 and 0.92, respectively.

Note that the linear HB structure predictor in (12) offers competitive performance at a lower complexity in comparison with BMARS and GPR that are complex data driven methods.

A linear predictor that is fitted on the pooled training set by least-squared-error achieves the average per-file PCC of 0.75 and average per-file RMSE of 0.64. This shows the effectiveness of considering heterogeneity via the linear HB structure, which allows the linear predictor to reach a comparable performance with more complex predictors.

### *B. Analysis of Task Configurations*

For different task configurations in Table I, we compute the mean and standard deviation of the per-file correlation coefficient and RMSE on the test sets. Fig. 2 illustrates the error-bar plot of the per-file performance of models trained using different configuration of tasks. The results indicate that different task configurations in Fig. 2 yield similar prediction performance.

We also compute the lower-bound $\mathcal{L}(q)$ for different task configurations and over all the training sets, and plot the means and the standard deviations in Fig. 3. Fig. 3 shows that configuration SD has the highest mean value of $\mathcal{L}(q)$. Thus, the Bayesian theory suggests that dividing the training data based on sex of the talker as well as database attributes results in the most plausible linear HB structure for Supplement 23 data set. The average value of RMSE and PCC metrics show that SD is not the best configuration. However, the confidence intervals are overlapping for all configurations. Such disagreements are often encountered when comparing the Bayesian log-model-evidence and RMSE and PCC metrics. We have discussed this issue in [37]. The Bayesian metric is directly influenced by the choice of priors, unlike RMSE and PCC, and is computed on the training set and not the test set.

We observe that the HB structure puts different weights on features in the task specific models.

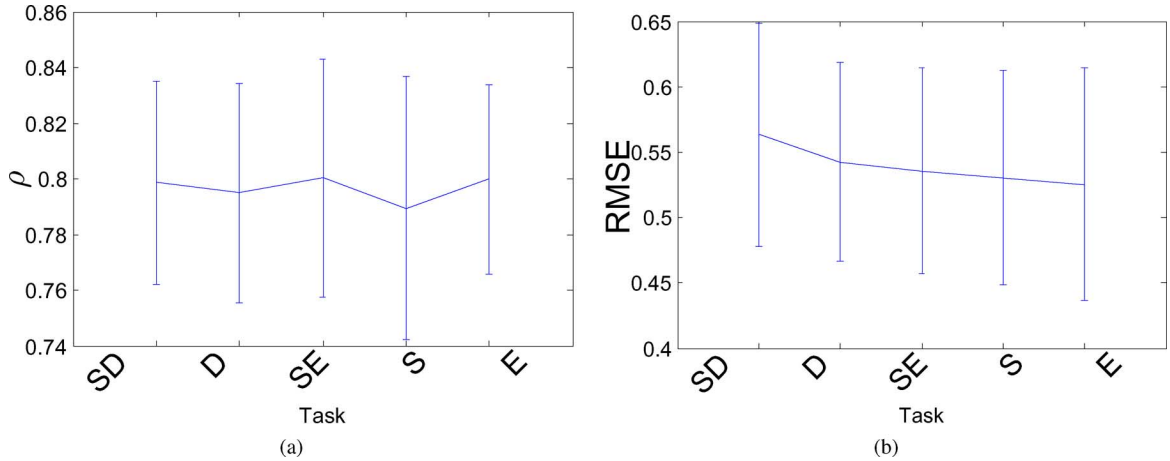



Fig. 2. Per-speech-file error-bar plots of PCC and RMSE for all task configurations. The horizontal axis corresponds to the task configuration: (Sex of the talker)-Database (SD), Database (D), (Sex of the talker)-Experiment (SE), Sex of the talker (S), and Experiment (E). (a) Error-bar plot of per-file PCC. (b) Error-bar plot of per-file RMSE.

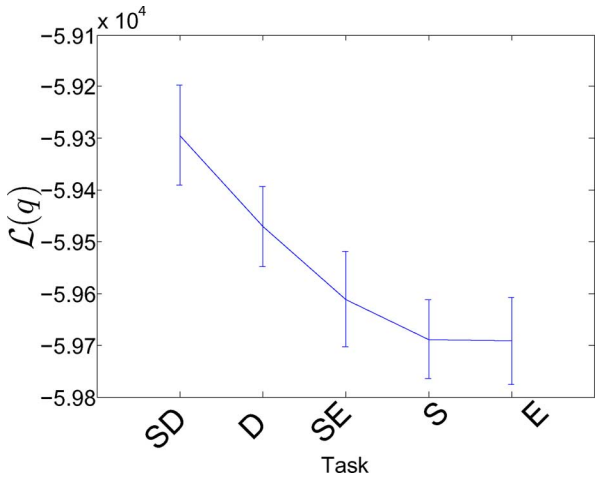



Fig. 3. Variational lower-bound to log model evidence for different task configurations. The horizontal axis corresponds to the model: (Sex of the talker)-Database (SD), Database (D), (Sex of the talker)-Experiment (SE), Sex of the talker (S), and Experiment (E). Larger numbers correspond to more plausible task configurations based on the given data.

## VII. Discussion

In our approach, attributes are used to divide the heterogeneous training data into tasks, and the predictor is a linear function that maps features into quality estimates. Thus, the complexity of our approach lies in the training phase instead of the deployment phase. This is in contrast with previous efforts to deploy machine learning algorithms for non-intrusive quality estimation, such as [24], [35], and [36], where the heterogeneous data is pooled into one set and important attributes such as language or laboratory are ignored.

To compare the performance of the Bayesian HB structure with the non-Bayesian P.563 , PCC, and RMSE were used. To compute these performance measures a point estimate is required instead of the predictive distribution in (3); thus, we use the mean of the predictive distribution $\Pr(y|\boldsymbol{\psi}, \mathcal{D}, \mathcal{M})$ as the speech quality estimate. Nevertheless, the predictive distribution is more informative than a point estimate. That is, the predictive distribution specifies not only the most probable quality estimate, but also the uncertainty about the estimate. Estimation uncertainty is not used in this work, but it can be most useful to perform active learning or Bayesian experimental design [43], which is not the scope of this work.

Public data sets for speech quality ratings are scarce: Supplement 23 is one of the few exceptions, which is often used in evaluation of narrow band speech quality measures. Because of its limited number of conditions and speech samples, this data set is usually augmented by other larger private data sets. Nevertheless, Supplement 23 contains substantial heterogeneity, which is confirmed by the performance gain of the predictor trained using the linear HB structure over other predictors of similar and higher complexity. Further experiments including larger data sets would provide more evidence; however, most speech quality data sets are not publicly available. We publish the details of our inference algorithm in Appendix A, which makes it possible to test our approach on other data sets as well.

The effectiveness of our approach is further confirmed by comparing our method and P.563. The complexity of our linear predictor is lower than the P.563 mapping, which was described in Section II. Furthermore P.563, due to limitations of the mapping, cannot reliably predict the quality of speech signals where both male and female talkers are present [41]. This limitation does not hold for our predictor which works for talkers of both sexes.

The complexity of implementing the feature mapping using our model is significantly lower than other methods such as P.563: for a $D$-dimensional feature space, we need only to store $D$ coefficients (feature weights) and perform $D$ multiplications. P.563 uses approximately $6 + 1$ such linear maps for the six class-dependent linear predictors and the global features. More importantly, P.563 requires five binary classifiers. BMARS and GPR need to store the MCMC posterior samples and the training-set data points, respectively, which limits the applicability of these two algorithms. The most important observation is that once we address heterogeneity in the data during the training phase of QA measures, we can achieve satisfactory performance with relatively simple structures for mapping the features into quality scores.

The complexity of Bayesian inference in our variation algorithm is substantially smaller than widely adopted Markov-chain Monte-Carlo (MCMC) sampling methods: our inference algorithm is an iterative process that converges typically in as

few as 20 iterations. Most of the complexity in each iteration lies in $m+1$ inversion of $D \times D$ matrices, where $m$ is the number of tasks and $D$ is the dimensionality of our feature-set. In our work, $D = 43$.

## VIII. Conclusion

Speech QA measures are generally trained on heterogeneous data sets. Heterogeneity is often neglected by pooling the data of various types into one large training set. Thus, QA measures, and in particular non-intrusive measures, require complex feature mappings to deal with the variability of statistical properties in the heterogeneous data. We demonstrated the promise of multi-task learning in dealing with data heterogeneity in Supplement 23 data set. Supplement 23 data set comprises different sources of heterogeneity including male and female speakers, different laboratories and languages and various distortion conditions. We realized the multi-task learning via a fully Bayesian linear HB structure, which outperforms other models such as BMARS and GPR trained on the pooled data. Our results were also competitive with the ITU-T P.563 standard.

We presented a fully Bayesian linear HB structure for non-intrusive speech quality assessment, where the training is carried out by an efficient variational inference algorithm. The inference algorithm approximates the posterior as well as the log-model-evidence. The details of the inference algorithm is given, which allows for easy implementation of our approach. The HB structure requires the heterogeneous training data to be divided into tasks. Tasks are defined using data attributes such as the origin of the data, the language, or sex of the talker, etc. The choice of optimal task configuration was formulated as a Bayesian model selection problem, where the Bayesian theory provides a formal tool, i.e., the model-evidence for selection of task configuration. We showed that while prediction-based measures such as Pearson correlation-coefficient and root-mean-squared error do not manifest which sets of tasks are better, the Bayesian model-evidence favors task sets that capture the heterogeneity in a more detailed manner.

By addressing the heterogeneity using the linear HB structure, we were able to create a simple linear predictor that performs better than other more complex methods for mapping features. In comparison, BMARS prediction can be done rapidly, but is memory intensive as samples of the posterior needs to be saved into memory. GPR also needs to store the training set which can be prohibitive if implementation on mobile devices are necessary or due to confidentiality of data sets. Sparse methods such as [44] exist for the GPR, which requires a sub-set of the training data to be stored at the cost of decreasing accuracy.

## Appendix A
## Bayesian Inference for the Linear HB Structure

Speech quality data sets are typically private. We give the details of our linear HB implementation, which allows for other researchers in the community to test it on their data sets. Here we provide an efficient variational method for the linear HB structure presented in Section IV-D. Our inference algorithm allows for approximating the log-model-evidence, which is used for selecting the task configuration. Computing this bound requires insignificant extra computations. While linear HB structures have been available, parameter estimation is typically done in maximum *a-posteriori* (MAP) method which may be prone to over-fitting and more importantly does not compute log-model-evidence.

*Variational Inference:* In Section IV-E, we defined the lower bound $\mathcal{L}(q)$ to the log-model-evidence, where $q(\boldsymbol{\theta})$ is our approximation to the posterior. To approximate the posterior, we start by assuming that $q(\boldsymbol{\theta})$ factorizes over the following partitioning of variables in $\boldsymbol{\theta}$:

$$q(\boldsymbol{\theta}) = q(\alpha)q(\lambda)q(W,\Lambda)\prod_{l} q(\boldsymbol{\omega}_l) \tag{16}$$

which assumes $\boldsymbol{\omega}_l$, $\{W,\Lambda\}$, $\lambda$, and $\alpha$ are independent in the posterior. This assumption on the posterior allows us to develop an iterative optimization algorithm for finding the $q^\star(\boldsymbol{\theta})$ that maximizes $\mathcal{L}(q)$ in (14). Maximization of the variational lower-bound amounts to minimization of the KL divergence, i.e., finding the closest distribution $q(\boldsymbol{\theta})$ under condition (16) to the posterior.

In each iteration of the algorithm one of the $q(\boldsymbol{\theta})$ factors is updated while the others are fixed. To update the $i$th factor we compute

$$\ln q^\star(\boldsymbol{\theta}_i) = \mathbb{E}_{j\neq i}[\ln \Pr(\boldsymbol{\theta}, \mathcal{D}|\mathcal{M})] + \text{const} \tag{17}$$

where $\mathbb{E}_{j\neq i}[]$ denotes the expectation with respect to variables present in the remaining factors $j \neq i$. That is, to update a certain factor, we compute the expectation of complete-data log-likelihood $\ln \Pr(\boldsymbol{\theta}, \mathcal{D}|\mathcal{M})$ with respect to remaining factors in $q(\boldsymbol{\theta})$.

The learning procedure starts by parameter initializing, which is covered in more detail as we describe the update equation for each factor. The derivations details of update equations are not included due to space limitations, but they are based on the factorized distribution we assumed in (16) and the update rule in (17). The complete-data log-likelihood is computed first as follows:

$$\ln \Pr(\boldsymbol{\theta}, \mathcal{D}, \mathcal{M}) = \ln \Pr(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) + \ln \Pr(\boldsymbol{\theta}|\mathcal{M}) \tag{18}$$

where

$$\ln \Pr(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) = \sum_{l=1}^{m}\sum_{i=1}^{n^l} \ln \mathcal{N}(y_{il}|\boldsymbol{\omega}_l^{\mathrm{T}}\boldsymbol{\psi}_{il}, \lambda^{-1}) \tag{19}$$

and

$$\begin{aligned}\ln \Pr(\boldsymbol{\theta}) = &+ \sum_{l=1}^{m} \ln \mathcal{N}(\boldsymbol{\omega}_l|W, (\lambda\Lambda)^{-1}) + \ln \mathcal{W}(\Lambda|\tau_0, \Sigma_0) \\ &+ \ln \mathcal{N}(W|W_0, (\alpha\Lambda)^{-1}) + \ln \text{Gamma}(\alpha|a_0^\alpha, b_0^\alpha) \\ &+ \ln \text{Gamma}(\lambda|a_0^\lambda, b_0^\lambda).\end{aligned} \tag{20}$$

Thus, we get

$$\begin{aligned}\ln \Pr(\boldsymbol{\theta}, \mathcal{D}|\mathcal{M}) = &+ (0.5(N+mD) + a_0^\lambda - 1)\ln\lambda - b_0^\lambda\lambda \\ &+ (a_0^\alpha - 1 + 0.5D)\ln\alpha - b_0^\alpha\alpha \\ &- 0.5\lambda\sum_{l=1}^{m} S_l + 0.5(\tau_0 + mD - 1)\ln|\Lambda| \\ &- 0.5\alpha(W-W_0)^{\mathrm{T}}\Lambda(W-W_0) \\ &- 0.5\text{trace}(\Sigma_0^{-1}\Lambda) + \text{const}\end{aligned} \tag{21}$$

where

$$S_l = (\boldsymbol{\omega}_l - W)^{\mathrm{T}}\Lambda(\boldsymbol{\omega}_l - W) + \sum_{i=1}^{n^l}(y_{il} - \boldsymbol{\omega}_l^{\mathrm{T}}\boldsymbol{\psi}_{il})^2 \quad (22)$$

and const is the logarithm of the normalization factors.

By substituting (21) in (17), we get the optimal factors for (16). The conjugate structure in the model prior results in optimal factors being the following members of the family of exponential distributions:

$$q^\star(\boldsymbol{\omega}_l) = \mathcal{N}(\boldsymbol{\omega}_l|\boldsymbol{\mu}_l, \mathbb{E}^{-1}[\lambda]\boldsymbol{V}_l) \quad (23)$$
$$q^\star(W|\Lambda) = \mathcal{N}(W|\boldsymbol{\mu}_W, \boldsymbol{V}_W) \quad (24)$$
$$q^\star(\Lambda) = \mathcal{W}(\Lambda|\tau_0 + m, \Sigma_N) \quad (25)$$
$$q^\star(\lambda) = \mathrm{Gamma}(\lambda|a_N^\lambda, b_N^\lambda) \quad (26)$$
$$q^\star(\alpha) = \mathrm{Gamma}(\alpha|a_N^\alpha, b_N^\alpha) \quad (27)$$

where

$$\boldsymbol{V}_l^{-1} = \sum_{i=1}^{n^l}\boldsymbol{\psi}_{il}\boldsymbol{\psi}_{il}^{\mathrm{T}} + \mathbb{E}[\Lambda] \quad (28)$$
$$\boldsymbol{\mu}_l = \boldsymbol{V}_l\left(\sum_{i=1}^{n^l}\boldsymbol{\psi}_{il}y_{il} + \mathbb{E}[\Lambda]\boldsymbol{\mu}_W\right) \quad (29)$$
$$\boldsymbol{\mu}_W = \kappa\left(\mathbb{E}[\lambda]\sum_{l=1}^{m}\boldsymbol{\mu}_l + \mathbb{E}[\alpha]W_0\right) \quad (30)$$
$$\boldsymbol{V}_W = \kappa\Lambda^{-1} \quad (31)$$
$$\Sigma_N^{-1} = +\Sigma_0^{-1} + \sum_{l=1}^{m}\left(\boldsymbol{V}_l + \mathbb{E}[\lambda]\boldsymbol{\mu}_l\boldsymbol{\mu}_l^{\mathrm{T}}\right) + \mathbb{E}[\alpha]W_0W_0^{\mathrm{T}} - \kappa^{-1}\boldsymbol{\mu}_W\boldsymbol{\mu}_W^{\mathrm{T}} \quad (32)$$
$$a_N^\lambda = a_0^\lambda + 0.5(N + Dm) \quad (33)$$
$$b_N^\lambda = b_0^\lambda + 0.5\sum_{l=1}^{m}\mathbb{E}[S_l] \quad (34)$$
$$a_N^\alpha = a_0^\alpha + 0.5D \quad (35)$$
$$b_N^\alpha = b_0^\alpha + 0.5(\boldsymbol{\mu}_W - W_0)^{\mathrm{T}}\mathbb{E}[\Lambda](\boldsymbol{\mu}_W - W_0) + 0.5\kappa D \quad (36)$$
$$\kappa^{-1} = m\mathbb{E}[\lambda] + \mathbb{E}[\alpha] \quad (37)$$

where $\mathbb{E}[\Lambda]$, $\mathbb{E}[\lambda]$, and $\mathbb{E}[\alpha]$ are computed according to distributions specified in (25) to (27) (Refer to, e.g., Appendix B, [40]). Equation (34), requires the following expectation:

$$\mathbb{E}[S_l] = +(\boldsymbol{\mu}_l - \boldsymbol{\mu}_W)^{\mathrm{T}}\mathbb{E}[\Lambda](\boldsymbol{\mu}_l - \boldsymbol{\mu}_W) + \sum_{i=1}^{n^l}(y_{il} - \boldsymbol{\mu}_l^{\mathrm{T}}\boldsymbol{\psi}_{il})^2 + \mathbb{E}^{-1}[\lambda]\sum_{i=1}^{n^l}\boldsymbol{\psi}_{il}^{\mathrm{T}}\boldsymbol{V}_l\boldsymbol{\psi}_{il} + \mathbb{E}^{-1}[\lambda]\mathrm{trace}(\boldsymbol{V}_l\Lambda) + \kappa D. \quad (38)$$

*Variational Lower-Bound:* The variational lower-bound $\mathcal{L}(q)$ defined in (14) is an approximation to the log-model-evidence $\Pr(\mathcal{D}|\mathcal{M})$ defined in (7). Task configurations are compared against each other by using $\mathcal{L}(q)$ to compare models $\mathcal{M}$.

Monotonic increase of the $\mathcal{L}(q)$ value through iterations offers a means for checking the correctness of the algorithm as well as convergence.

To compute the $\mathcal{L}(q)$ for the linear HB structure, first (14) is expresses as

$$\mathcal{L}(q) = \mathbb{E}[\ln\Pr(\boldsymbol{\theta}, \mathcal{D}|\mathcal{M})] + \mathrm{H}(q) \quad (39)$$

where $\mathrm{H}(q)$ is the entropy of $q^\star(\boldsymbol{\theta})$ distributions and

$$\mathbb{E}[\ln\Pr(\boldsymbol{\theta}, \mathcal{D}|\mathcal{M})] = +0.5(\tau_0 + mD - 1)\mathbb{E}[\ln|\Lambda|] - 0.5\mathrm{trace}(\Sigma_0^{-1}\mathbb{E}[\Lambda]) + \mathrm{const} + (a_N^\alpha - 1)\mathbb{E}[\ln\alpha] - a_N^\alpha + (a_N^\lambda - 1)\mathbb{E}[\ln\lambda] - a_N^\lambda \quad (40)$$

where const is the logarithm of the normalization factors. The expectations $\mathbb{E}[\ln\lambda]$, $\mathbb{E}[\ln\alpha]$, and $\mathbb{E}[\ln\Lambda]$ are computed according to distributions specified in (25) to (27)

## Appendix B
## BMARS and GPR

MARS and BMARS have been adopted before for speech quality estimation both for intrusive and non-intrusive feature sets as explained in Section I. We use identical choices as Petkov *et al.* [34] for setting up our BMARS model: additive BMARS model with step basis functions. We limit the maximum number of basis functions to 300. BMARS uses MCMC sampling, depending on the experiment at least we generate 40000 samples as burn-in period, and collect at least 40 000 samples afterwards. The difference between our work and [34] is the features extracted from the speech signal: we use P.563 features, while in [34] modulation spectrum features were used.

We use the GPR implementation of Rasmussen and Williams [45]. The crucial decision when deploying GPR is the choice of the *kernel*, i.e., the distance metric between feature vectors $\mathrm{d}(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)$. We compared nine different kernels for GPR and selected the best one. The kernels are: linear, linear with automatic relevance determination (ARD), Matern with two parameters settings $\nu = 3/2$, and $\nu = 5/2$, neural network, rational quadratic covariance function with ARD, isotropic rational quadratic covariance function, squared exponential covariance function with ARD and isotropic squared exponential covariance function. ARD is a mechanism that allows for feature selection [46]. The kernel descriptions are available in the Rasmussen and Williams book as well as the online documentation [45]. In our experiments, the best performance results were obtained when using the following kernels: Matern with $\nu = 3/2$ and $\nu = 5/2$, isotropic rational quadratic and isotropic squared exponential covariance functions.

## References

[1] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[2] Methods for Subjective Determination of Transmission Quality ITU, Geneva, Switzerland, 1996, ITU-T Rec. P.800.
[3] P. Kroon, W. B. Kleijn, and K. Paliwal, "Evaluation of speech coders," *Speech Coding Synth.*, pp. 467–494, 1995.

[4] "Mean opinion score (MOS) terminology," ITU, Geneva, Switzerland, 2003, ITU-T Rec. P.800.1.
[5] A. Rix, M. Hollier, A. Hekstra, and J. Beerends, "Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment part I-time-delay compensation," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp. 755–764, 2002.
[6] J. Beerends, A. Hekstra, A. Rix, and M. Hollier, "Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment part II-psychoacoustic model," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp. 765–778, 2002.
[7] "Single-ended method for objective speech quality assessment in narrow-band telephony," Applications ITU, Geneva, Switzerland, 2004, ITU-T Rec. P.563.
[8] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.
[9] J. Baxter, "A model of inductive bias learning," *J. Artif. Intell. Res.*, vol. 12, pp. 149–198, 2000.
[10] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*. London, U.K.: Chapman & Hall, 1995, Texts in Statistical Science.
[11] "ITU-T coded-speech database," ITU, Geneva, Switzerland, 1998, ITU-T Rec. P.Suppl. 23.
[12] D. Denison, B. Mallick, and A. Smith, "Bayesian MARS," *Statist. Comput.*, vol. 8, no. 4, pp. 337–346, 1998.
[13] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.
[14] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, May 1982, vol. 7, pp. 1278–1281.
[15] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Sel. Areas Commun.*, vol. 10, no. 5, pp. 819–829, Jun. 1992.
[16] M. Hollier, M. Hawksford, and D. Guard, "Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain," *IEE Proc.-Vision Image Signal Process.*, vol. 141, no. 3, pp. 203–208, Jun. 1994.
[17] J. Beerends and J. Stemerdink, "A perceptual speech-quality measure based on a psychoacoustic sound representation," *J. Audio Eng. Soc.*, vol. 42, no. 3, pp. 115–123, 1994.
[18] S. Voran, "Objective estimation of perceived speech quality Part I: Development of the measuring normalizing block technique," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 4, pp. 371–382, Jul. 1999.
[19] S. Voran, "Objective estimation of perceived speech quality Part II: Evaluation of the measuring normalizing block technique," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 4, pp. 383–390, Jul. 1999.
[20] J. Liang and R. Kubichek, *Output-Based Objective Speech Quality*, vol. 3, pp. 1719–1723, Jun. 1994.
[21] P. Gray, M. Hollier, and R. Massara, "Non-intrusive speech-quality assessment using vocal-tract models," *IEE Proc.-Vis. Image Signal Process.*, vol. 147, no. 6, pp. 493–501, Dec. 2000.
[22] D. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pt. 2, pp. 821–831, Sep. 2005.
[23] D. Kim and M. Tarraf, *Enhanced Perceptual Model for Non-Intrusive Speech Quality Assessment*, vol. 1, pp. 829–832, May 2006.
[24] T. Falk and W. Chan, "Nonintrusive speech quality estimation using Gaussian mixture models," *IEEE Signal Process. Lett.*, vol. 13, no. 2, p. 108, 2006.
[25] T. Falk and W. Chan, "Single-ended speech quality measurement using machine learning methods," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1935–1947, Nov. 2006.
[26] T. Falk, H. Yuan, and W. Chan, "Single-ended quality measurement of noise suppressed speech based on Kullback–Leibler distances," *J. Multimedia*, vol. 2, no. 5, p. 19, 2007.
[27] G. Chen and V. Parsa, "Bayesian model based non-intrusive speech quality evaluation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, vol. 1, pp. 385–388.
[28] A. Rix and P. Gray, "NiQA-Non-intrusive speech quality assessment," *Contribution UIT-T COM*, 2001.
[29] "NiQA-Product Description," Tech. Rep. Psytechnics Limited, Jan. 2003. [Online]. Available: http://www.psytechnics.com/pages/products/niqa.php
[30] "NiNA – SwissQual's Non-Intrusive Algorithm for Estimating the Subjective Quality of Live Speech," Tech. Rep. SwissQual Incorporated, Jun. 2001. [Online]. Available: http://www.swissqual.com/NiNA-module.aspx
[31] J. Friedman, "Multivariate adaptive regression splines," *Ann. Statist.*, vol. 19, no. 1, pp. 1–67, 1991.
[32] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, p. 229, Jan. 2008.
[33] W. Zha and W. Chan, "Objective speech quality measurement using statistical data mining," *EURASIP J. Appl. Signal Process.*, vol. 2005, p. 1424, 2005.
[34] P. N. Petkov, S. I. Mossavat, and W. B. Kleijn, "A Bayesian approach to non-intrusive quality assessment of speech," in *Proc. Int. Conf. Spoken Lang. Process.*, Brighton, U.K., 2009, pp. 2875–2878.
[35] D. Picovici and A. Mahdi, "Output-based objective speech quality measure using self-organizing map," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2003, vol. 1, pp. 476–479.
[36] M. Narwaria, W. Lin, I. McLoughlin, S. Emmanuel, and C. Tien, "Non-intrusive speech quality assessment with support vector regression," in *Advances in Multimedia Modeling*. Berlin/Heidelberg, Germany: Springer, Lecture Notes in Computer Science, pp. 325–335.
[37] S. I. Mossavat, O. Amft, B. de Vries, P. N. Petkov, and W. B. Kleijn, "A Bayesian hierarchical mixture of experts approach to estimate speech quality," in *Proc. Int. Workshop Quality of Multimedia Experience*, 2010.
[38] C. Bishop and M. Svensén, "Bayesian hierarchical mixtures of experts," in *Proc. 19th Conf. Uncertainty Artif. Intell.*, 2003, pp. 57–64.
[39] K. Yu, V. Tresp, and A. Schwaighofer, "Learning Gaussian processes from multiple tasks," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, p. 1019.
[40] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
[41] L. Malfait, J. Berger, and M. Kastner, "P. 563-the ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 6, pp. 1924–1934, Nov. 2006.
[42] "Subjective test plan for characterization of an 8 kbit/s speech codec," ITU-T Study Group 12, 1995, ITU-T SQ-46.95R3.
[43] D. Cohn, Z. Ghahramani, and M. Jordan, "Active learning with statistical models," *Arxiv Preprint cs/9603104*, 1996.
[44] J. Quiñonero-Candela and C. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," *J. Mach. Learn. Res.*, vol. 6, pp. 1939–1959, 2005.
[45] C. Rasmussen and C. Williams, "Gaussian process regression and classification," [Online]. Available: http://www.gaussianprocess.org/gpml/
[46] D. MacKay, "A practical Bayesian framework for backpropagation networks," *Neural Comput.*, vol. 4, no. 3, pp. 448–472, 1992.

**Iman Mossavat** (S'09) was born in Mashhad, Iran, in 1979. He received the B.Sc. degree from the Department of Electrical Engineering, Ferdowsi University of Mashhad, in 2002, the M.Sc. degree from the Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran, in 2005, and the M.Eng. degree from the Department of Electrical and Computer Engineering, National University of Singapore, in 2008. He is currently working toward the Ph.D. degree in electrical engineering with the Eindhoven University of Technology, Eindhoven, The Netherlands.

His research interests include signal processing and Bayesian machine learning.

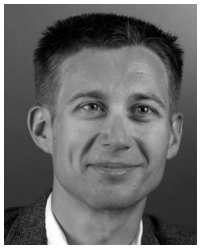
**Petko N. Petkov** (S'10) was born in Sofia, Bulgaria, in 1980. He received the B.Sc. degree in communication engineering from the Technical University of Sofia and the M.Sc. degree in electrical engineering from the Royal Institute of Technology (KTH), Stockholm, Sweden. He is currently pursuing the Ph.D. degree at the Sound and Image Processing Lab, School of Electrical Engineering, KTH.

He was a Research and Development Engineer with Global IP Sound in 2006–2007. He has held visiting researcher positions at the Tampere University of Technology and KTH. His research interests include the design of applied statistical regression and classification models and the application of control theory to problems in audio and video processing.

**W. Bastiaan Kleijn** (M'88–SM'97–F'99) received the M.S. degree in electrical engineering from Stanford University, Stanford, CA, the M.S. degree in physics and the Ph.D. degree in soil science, both from the University of California, Riverside, and the Ph.D. degree in electrical engineering from the Delft University of Technology, Delft, The Netherlands.

He has been a Professor of Electronic Engineering at Victoria University of Wellington, Wellington, New Zealand, since 2010. He is also a Professor at the School of Electrical Engineering at KTH (the Royal Institute of Technology), Stockholm, Sweden, where he was until recently Head of the Sound and Image Processing Laboratory. He worked on speech processing at AT&T Bell Laboratories from 1984 to 1996. He was a founder of Global IP Solutions, which was acquired by Google in 2010.

Dr. Kleijn is on the Editorial Board of *Signal Processing* and has been on the Boards of the IEEE TRANSACTIONS OF SPEECH AND AUDIO PROCESSING, IEEE SIGNAL PROCESSING LETTERS, *IEEE Signal Processing Magazine*, and the *EURASIP Journal of Applied Signal Processing*. He has been a member of several IEEE technical committees, and a Technical Chair of EUSIPCO 2010, ICASSP'99, the 1997 and 1999 IEEE Speech Coding Workshops, and a General Chair of the 1999 IEEE Signal Processing for Multimedia Workshop.



**Oliver Amft** (M'08) received the Dipl.-Ing. (M.Sc.) degree from Chemnitz Technical University, Chemnitz, Germany, in 1999 and the Dr. sc. ETH (Ph.D.) degree from ETH Zurich, Zurich, Switzerland, in 2008.

Until 2004, he was an R&D Project Manager with ABB Inc., leading product developments in embedded communication systems for five years. He is currently an Assistant Professor Signal Processing Systems Group, Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands, leading the ACTLab research group. He is also a Senior Research advisor at the Wearable Computing Lab, ETH Zurich. His research focuses on multi-modal activity recognition and human behavior inference algorithms with applications in healthcare, wellness, sports, and smart buildings.

Dr. Amft received the EWSN/CONET award in 2009 for his Ph.D. thesis.