SPARSE BAYESIAN HIERARCHICAL MIXTURE OF EXPERTS

Iman Mossavat, Oliver Amft

Signal Processing Systems Group, Department of Electrical Engineering
Eindhoven University of Technology, The Netherlands

ABSTRACT

Hierarchical mixture of experts (HME) is a widely adopted probabilistic divide-and-conquer regression model. We extend the variational inference algorithm for HME by using automatic relevance determination (ARD) priors. Unlike Gaussian priors, ARD allows for a few model parameters to take on large values, while forcing others to zero. Thus, using ARD priors encourages sparse models. Sparsity is known to be advantageous to the generalization capability as well as interpretability of the models. We present the variational inference algorithm for sparse HME in detail. Subsequently, we evaluate the sparse HME approach in building objective speech quality assessment algorithms, that are required to determine the quality of service in telecommunication networks.

Index Terms— Mixture, sparse, variational, inference, Bayesian, speech quality.

1. INTRODUCTION

Hierarchical mixture of experts (HME) is a probabilistic divide-and-conquer regression method [1]. HME divides the feature space into regions, and assigns a linear regressor to each region. HME is a conditional mixture of linear regressors. Mixture components are called *experts*. Experts are combined in the mixture using weights, called mixing coefficients. Mixing coefficients depend on the input feature vector, and are determined by the *gating network*: a tree structure with binary classifiers, or *gates*, at its internal nodes. The gating network task is to divide the feature space into regions by adapting the mixing coefficients based on the feature vector.

Since HME is a probabilistic method, decision boundaries among the regions are 'soft'. Thus a given data point may lie simultaneously in several regions. The following advantages make HME more favorable than other divide-and-conquer regression methods with 'hard' decision boundaries such as CART [2]: the capability to model multi-modal distributions as explained in [3], less abrupt fluctuations near decision boundaries [4], and smaller variance as described in [1].

CART models allow easier interpretation, which is an advantage over HME. CART divides the space by specifying thresholds on certain feature values, whereas HME uses probabilistic binary classifiers in its gating network. Existing inference algorithms for HME models, e.g. [1, 3] do not yield sparse results, making HME models hard to interpret.

In this work, we extend the variational Bayesian HME of Bishop and Svensén [3], in which they approximate the posterior using the *mean-field* method [4]. We specify new priors for HME based on the automatic relevance determination (ARD) mechanism [5, 6]. ARD allows for experts and gates to be sparse. Sparsity often promotes interpretability as well as better generalization. We present the details of the approximate inference algorithm for the sparse HME.

The new sparse HME is subsequently evaluated by developing a non-intrusive objective speech quality assessment (QA) algorithm, which is an estimator of the perceived speech quality. The state of the art algorithm in non-intrusive speech QA is represented by the ITU-T recommendation P.563 [7], which exploits a divide-and-conquer strategy. P.563 classifies speech signals transmitted over a telecommunication network based on the talker sex as well as how the network distorted the speech, e.g. high-level background noise, speech robotization, etc. P.563 uses hard binary classifiers, thus its output changes abruptly near distortion class boundaries. P.563 is limited to assigning only one class to each signal, thus it cannot reliably predict the quality when both male and female talkers are present in the speech signal [7]. In [8], we use the variational Bayesian HME of [3] to model the speech quality data in the ITU-T coded-speech data-set, Supplement 23 [9]. HME soft decision boundaries effectively addresses the non-smoothness limitation of P.563, and achieves competitive prediction performance with P.563. However, even with two experts, the HME model is hard to interpret.

2. HME MODEL

HME defines a mixture distribution over y, the target variable, conditioned on the D-dimensional feature vector $\boldsymbol{\psi}$, which is

$$\Pr(y|\boldsymbol{\psi},\boldsymbol{\theta}) \quad = \quad \sum_{k=1}^{C} g_k(\boldsymbol{\psi})\mathcal{N}(y|\boldsymbol{\omega}_k^{\mathrm{T}}\boldsymbol{\psi},\tau_k^{-1}), \qquad (1)$$

where $\boldsymbol{\theta}$ denote the model parameters, C denotes the number of experts, and $g_k(\boldsymbol{\psi})$ denotes the mixing-coefficient for the k-th

Iman Mossavat is funded by STW project 07605: Personalization of Hearing Aids Through Bayesian Preference Elicitation (HearClip).

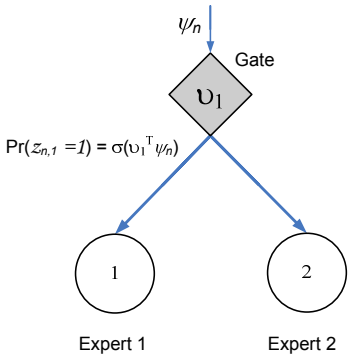

Fig. 1. The gating network of a HME model with one gate and two experts. The gate is a binary classifier. For n-th input $\boldsymbol{\psi}_n$, the latent binary variable $z_{n,1}$ indicates the gate output, with $z_{n,1} = 1$ corresponding to selecting the expert on the left. The gate selects the output branch on its left by probability $\sigma(\boldsymbol{\upsilon}_1^{\mathrm{T}}\boldsymbol{\psi}_n)$, where $\sigma(.)$ is the logistic function. The mixing-coefficient for each expert is the product of the link probabilities on the path that connects the root node to the expert.

expert. Experts correspond to Gaussian distributions, denoted by $\mathcal{N}$. The mean of the k-th expert is given by $\boldsymbol{\omega}_k^{\mathrm{T}}\boldsymbol{\psi}$ and its precision (inverse of variance) by τ_k. The gating network determines $g_k(\boldsymbol{\psi})$ as described in the following section.

Suppose the HME model has C experts and M gates. From this point forward, we use $k = 1, \ldots, C$, $l = 1, \ldots, M$, and $n = 1, \cdots, N$ to denote respectively the expert, the gate, and the data point indices. In addition, the i-th element of vector $\boldsymbol{x}$ is denoted by $\boldsymbol{x}(i)$.

2.1. Gating Network

We define the latent variable $z_{n,l} \in \{0, 1\}$, corresponding to the l-th gate and n-th data point $\boldsymbol{\psi}_n$, such that $z_{n,l} = 1$ indicates the l-th gate left-side branch is chosen for $\boldsymbol{\psi}_n$. The mixing-coefficients in Equation (1) are defined as

$$g_k(\boldsymbol{\psi}_n) = \Pr(\zeta_{n,k} = 1), \tag{2}$$

where

$$\zeta_{n,k} = \prod_{l=1}^{M} z_{n,l}^{\boldsymbol{S}^{\mathrm{L}}(k,l)} \; (1 - z_{n,l})^{\boldsymbol{S}^{\mathrm{R}}(k,l)}, \tag{3}$$

where the gating network topology is specified by binary matrices $\boldsymbol{S}^L$ and $\boldsymbol{S}^R$, where $\boldsymbol{S}^{\mathrm{L}}(k, l) = 1$ if k-th expert is on the 'left' sub-tree of l-th gate, and zero otherwise. Similarly, $\boldsymbol{S}^{\mathrm{R}}(k, l) = 1$ if k-th expert is on the 'right' sub-tree of l-th gate, and zero otherwise.

It is straightforward to verify that $\zeta_{n,k} = 1$, if and only if, for the point $\boldsymbol{\psi}_n$ the gates connect the root to the k-th expert. The probability distribution of $z_{n,l}$ is

$$\Pr(z_{n,l}|\boldsymbol{\psi}_n, \boldsymbol{\upsilon}_l) = \sigma(\boldsymbol{\upsilon}_l^{\mathrm{T}}\boldsymbol{\psi}_n)^{z_{n,l}}(1 - \sigma(\boldsymbol{\upsilon}_l^{\mathrm{T}}\boldsymbol{\psi}_n))^{(1-z_{n,l})}, \tag{4}$$

where $\sigma(x) = 1/\,(1 + \exp(-x))$ is the sigmoid function, and $\boldsymbol{\upsilon}_l$ is the weight vector of l-th gate. We illustrate the gating network of a HME model with two experts in Figure 1.

3. ARD PRIOR

To infer sparse experts and gates parameters from the data, we incorporate ARD in the priors for $\boldsymbol{\omega}_k$ and $\boldsymbol{\upsilon}_l$. The hierarchical prior for the l-th gate is defined as

$$\Pr(\boldsymbol{\upsilon}_l|\boldsymbol{\beta}_l) = \prod_{i=1}^{D} \mathcal{N}(\boldsymbol{\upsilon}_l(i)|0, \boldsymbol{\beta}_l^{-1}(i)) \tag{5}$$

$$\Pr(\boldsymbol{\beta}_l) = \prod_{i=1}^{D} \mathrm{Gamma}(\boldsymbol{\beta}_l(i)|a_0, b_0), \tag{6}$$

where D is the dimensionality of the feature vector $\boldsymbol{\psi}$ and Gamma denotes the Gamma distribution. By $\boldsymbol{\beta}_l(i)$ we denote the precision of the i-th feature weight, i.e., $\boldsymbol{\upsilon}_l(i)$. A large value for $\boldsymbol{\beta}_l(i)$ results in the $\boldsymbol{\upsilon}_l(i)$ prior to be concentrated around zero, thus forcing $\boldsymbol{\upsilon}_l(i)$ to small values. Learning $\boldsymbol{\beta}_l(i)$ along with $\boldsymbol{\upsilon}_l(i)$ results in determination of feature relevance in the l-th gate and encourages sparsity ([4], Chapter 7). Similarly, the k-th expert prior is

$$\Pr(\boldsymbol{\omega}_k|\tau_k, \boldsymbol{\alpha}_k) = \prod_{i=1}^{D} \mathcal{N}(\boldsymbol{\omega}_k(i)|0, \tau_k^{-1}\boldsymbol{\alpha}_k^{-1}(i)) \tag{7}$$

$$\Pr(\tau_k|a_0, b_0) = \mathrm{Gamma}(\tau_k|a_0, b_0) \tag{8}$$

$$\Pr(\boldsymbol{\alpha}_k|a_0, b_0) = \prod_{i=1}^{D} \mathrm{Gamma}(\boldsymbol{\alpha}_k(i)|a_0, b_0), \tag{9}$$

where $\boldsymbol{\alpha}_k(i)$ is the shrinkage parameter corresponding to $\boldsymbol{\omega}_k(i)$. Jointly inferring $\boldsymbol{\alpha}_k(i)$ and $\boldsymbol{\omega}_k(i)$ results in determination of feature relevance for the k-th expert.

4. VARIATIONAL INFERENCE

For non-sparse HME, an efficient Bayesian inference algorithm, based on mean-field approximation method, is given in [3]. We first review the mean-field approximation briefly [4], and then proceed to present our inference algorithm that is working with priors specified in the previous section. As in [3], we approximate the sigmoid function $\sigma(x)$ such that the conjugacy of the model is restored.

By $\boldsymbol{\theta}$, $\mathcal{D}$, and $\mathcal{M}$ we denote the model parameters, the data set and the model, respectively. In HME, the model $\mathcal{M}$ specifies the gating network topology and the hyper-priors. To approximate the posterior $\Pr(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})$, we start by writing the log model evidence $\ln \Pr(\mathcal{D}|\mathcal{M})$ as

$$\ln \Pr(\mathcal{D}|\mathcal{M}) = \mathcal{L}(q) + \mathrm{KL}(q\|p) \tag{10}$$

$$\mathcal{L}(q) = \int q(\boldsymbol{\theta}) \ln \frac{\Pr(\boldsymbol{\theta}, \mathcal{D}|\mathcal{M})}{\Pr(\boldsymbol{\theta}|\mathcal{M})} \mathrm{d}\boldsymbol{\theta}, \tag{11}$$

where $q(\boldsymbol{\theta})$ denotes an arbitrary distribution. By $\mathrm{KL}(q||p)$ we denote the Kullback-Leibler (KL) divergence between distributions $q(\boldsymbol{\theta})$ and the posterior $\Pr(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$. KL divergence is always non-negative, thus $\mathcal{L}(q) \leq \ln \Pr(\mathcal{D}|\mathcal{M})$, and $\mathcal{L}(q)$ is a variational *lower-bound* to log-model-evidence. The goal here is to develop a computationally tractable algorithm, which accurately approximates the posterior by $q(\boldsymbol{\theta})$. For sparse HME an efficient iterative algorithm for finding the optimal $q(\boldsymbol{\theta})$ exists, once we 'assume' that $q(\boldsymbol{\theta})$ factorizes as

$$q(\boldsymbol{\theta}) = \prod_{k=1}^{C} q(\boldsymbol{\omega}_k, \tau_k) q(\boldsymbol{\alpha}_k) \prod_{l=1}^{M} q(\boldsymbol{v}_l) q(\boldsymbol{\beta}_l) \prod_{n=1}^{N} q(z_{n,l}). \quad (12)$$

Equation (12) specifies a set of distributions. It is possible to efficiently find the member of this set, $q^\star(\boldsymbol{\theta})$, which maximizes the lower-bound $\mathcal{L}(q^\star)$. The assumed factorization in Equation (12) makes efficient optimization of $\mathcal{L}(q)$ possible. In each iteration of the $\mathcal{L}(q)$ optimization, one of the factors is updated while other factors are constant. Given a conjugate-exponential structure, update equations have closed form formulae. We present the update equations, but the formula for $\mathcal{L}(q)$ and the derivations are omitted due to space limitation.

4.1. Update Equations

Computing the update equations for the gates are complicated by the fact that the conjugate-exponential structure of the HME model is thwarted by the sigmoid function in Equation (4). We use the parametric approximation of the sigmoid function in [3]. For each gate l and data point $\boldsymbol{\psi}_n$, we define a variational parameter $\gamma_{n,l}$. In each iteration we update $\gamma_{n,l}$ along with the other factors in Equation (12). In the following we use $i = 1, \cdots, D$ to denote the feature index. The terms of optimal $q^\star(\boldsymbol{\theta})$ in Equation (12) are

$$\begin{aligned}
q^\star(z_{n,l}) &= \sigma(h_{n,l})^{z_{n,l}} (1-\sigma(h_{n,l}))^{1-z_{n,l}} && (13)\\
q^\star(\boldsymbol{\omega}_k|\tau_k) &= \mathcal{N}(\boldsymbol{\omega}_k|\bar{\boldsymbol{\omega}_k}, \tau_k^{-1}\mathbf{V}_k) && (14)\\
q^\star(\tau_k) &= \mathrm{Gamma}(\tau_k|a_{\tau_k}, b_{\tau_k}) && (15)\\
q^\star(\boldsymbol{v}_l) &= \mathcal{N}(\boldsymbol{v}_l|\bar{\boldsymbol{v}_l}, \boldsymbol{\Lambda}_l) && (16)\\
q^\star(\boldsymbol{\alpha}_k) &= \prod_{i=1}^{D} \mathrm{Gamma}(\boldsymbol{\alpha}_k(i)|\boldsymbol{a}_{\boldsymbol{\alpha}_k}(i), \boldsymbol{b}_{\boldsymbol{\alpha}_k}(i)) && (17)\\
q^\star(\boldsymbol{\beta}_l) &= \prod_{i=1}^{D} \mathrm{Gamma}(\boldsymbol{\beta}_l(i)|\boldsymbol{a}_{\boldsymbol{\beta}_l}(i), \boldsymbol{b}_{\boldsymbol{\beta}_l}(i)), && (18)
\end{aligned}$$

where the update equations are

$$\begin{aligned}
\gamma_{n,l} &= \boldsymbol{\psi}_n^{\mathrm{T}} \boldsymbol{\Lambda}_l \boldsymbol{\psi}_n + (\bar{\boldsymbol{v}_l}^{\mathrm{T}} \boldsymbol{\psi}_n)^2 && (19)\\
h_{n,l} &= \sum_{k \in \mathcal{E}_l^{\mathrm{R}}} \zeta'_{n,k} \mathbb{E}\big[\log \mathcal{N}(y_n|\boldsymbol{\omega}_k^{\mathrm{T}}\boldsymbol{\psi}_n, \tau_k^{-1})\big] - \cdots \\
&\quad \sum_{k \in \mathcal{E}_l^{\mathrm{L}}} \zeta'_{n,k} \mathbb{E}\big[\log \mathcal{N}(y_n|\boldsymbol{\omega}_k^{\mathrm{T}}\boldsymbol{\psi}_n, \tau_k^{-1})\big] && (20)\\
\mathbf{V}_k^{-1} &= \mathbb{E}[\mathbf{A}_k] + \sum_{n=1}^{N} \mathbb{E}[\zeta_{n,k}] \boldsymbol{\psi}_n \boldsymbol{\psi}_n^{\mathrm{T}} && (21)\\
\bar{\boldsymbol{\omega}_k} &= \mathbf{V}_k \sum_{n=1}^{N} \mathbb{E}[\zeta_{n,k}] y_n \boldsymbol{\psi}_n && (22)\\
a_{\tau_k} &= a_0 + 0.5 \sum_{n=1}^{N} \mathbb{E}[\zeta_{n,k}] && (23)\\
b_{\tau_k} &= b_0 + 0.5 \left(\sum_{n=1}^{N} \mathbb{E}[\zeta_{n,k}] \, (\bar{\boldsymbol{\omega}_k} \boldsymbol{\psi}_n - y_n)^2 + \cdots \right. \\
&\quad \left. \bar{\boldsymbol{\omega}_k}^{\mathrm{T}} \mathbb{E}[\mathbf{A}_k] \bar{\boldsymbol{\omega}_k} \right) && (24)\\
\boldsymbol{\Lambda}_l^{-1} &= \mathbb{E}[\mathbf{B}_l] + 2 \sum_{n=1}^{N} \lambda(\gamma_{n,l}) \boldsymbol{\psi}_n \boldsymbol{\psi}_n^{\mathrm{T}} && (25)\\
\bar{\boldsymbol{v}_l} &= \boldsymbol{\Lambda}_l \sum_{n=1}^{N} (\mathbb{E}[z_{n,l}] - 0.5) \boldsymbol{\psi}_n && (26)\\
\boldsymbol{a}_{\boldsymbol{\alpha}_k}(i) &= a_0 + 0.5 && (27)\\
\boldsymbol{b}_{\boldsymbol{\alpha}_k}(i) &= b_0 + 0.5 \mathbb{E}[\tau_k] \bar{\boldsymbol{\omega}_k}(i)^2 + \mathbf{V}_k(i,i) && (28)\\
\boldsymbol{a}_{\boldsymbol{\beta}_l}(i) &= a_0 + 0.5 && (29)\\
\boldsymbol{b}_{\boldsymbol{\beta}_l}(i) &= b_0 + 0.5 \bar{\boldsymbol{v}_l}(i)^2, && (30)
\end{aligned}$$

where $\mathbb{E}$ denotes the expectation with respect to $q^\star(\boldsymbol{\theta})$, $\mathcal{E}_l^{\mathrm{R}}$ and $\mathcal{E}_l^{\mathrm{L}}$ denote the set of experts on the right-hand-side and left-hand-side of the l-th gate respectively, and $\zeta'_{n,k}$ is computed as $\zeta_{n,k}$ in Equation (3) with terms corresponding to l-th gate being omitted. By $\mathbf{B}_l$ and $\mathbf{A}_k$ we denote diagonal matrices with $\boldsymbol{\beta}_l$ and $\boldsymbol{\alpha}_k$ on their diagonal, respectively. The i-th element on the diagonal of $\mathbf{V}_k$ is denoted by $\mathbf{V}_k(i,i)$. Finally, the expectation of $\mathbb{E}[\zeta_{n,k}]$ is computed as

$$\mathbb{E}[\zeta_{n,k}] = \prod_{l=1}^{M} \sigma(h_{n,l})^{\boldsymbol{S}^{\mathrm{L}}(k,l)} \left(1 - \sigma(h_{n,l})\right)^{\boldsymbol{S}^{\mathrm{R}}(k,l)} \quad (31)$$

Considering that $q^\star(\tau_k)$, $q^\star(\boldsymbol{\alpha}_k)$, and $q^\star(\boldsymbol{\beta}_l)$ are Gamma distributed, we can readily compute the expectation involving variables τ_k, $\boldsymbol{\alpha}_k$, and $\boldsymbol{\beta}_l$.

ARD is reflected in Equations (28) and (30), where $\boldsymbol{b}_{\boldsymbol{\alpha}_k}(i)$ and $\boldsymbol{b}_{\boldsymbol{\beta}_l}(i)$ specify the prior over the i-th feature in the k-th expert, and the l-th gate, respectively.

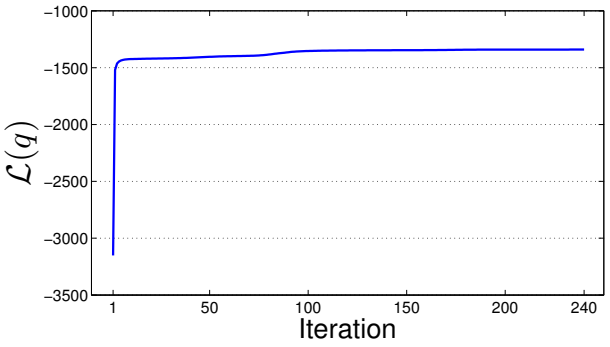

Fig. 2. Convergence of lower-bound to log-model-evidence, $\mathcal{L}(q)$, for random initialization.

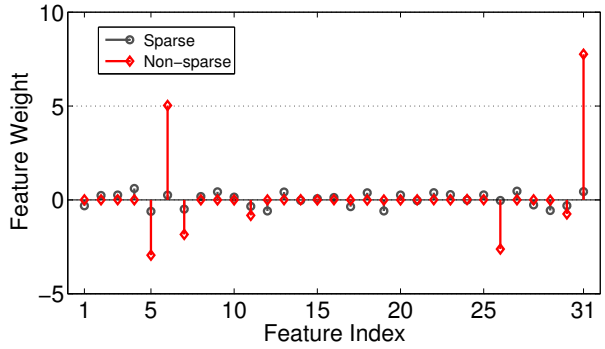

Fig. 3. Feature weights of the only gate in the network for non-sparse and sparse HME models. The features are normalized to zero-mean and unit variance.

5. EXPERIMENTAL EVALUATION

To evaluate the effectiveness of sparse HME, we compared its performance against P.563 as well as the variational Bayesian HME of [3]. We used the ITU-T coded-speech data-set, Supplement 23 [9]. More details about our data-set and experimental method is given in [8]. A measure called the condition-averaged correlation-coefficient (CACC) is often used to quantify the accuracy of quality assessment algorithms in cross-validation experiments. For brevity we omit details of the experiments, which are identical to [8]. The CACC performance of the new sparse HME matched that of the non-sparse HME, both yielding 0.88. This result confirms an improvement compared to the P.563, which yielded 0.87. We used Gamma hyper-parameters $a_0 = 10^{-2}$ and $b_0 = 10^{-4}$ to obtain broad support for the priors.

Figure 2 shows the convergence of the lower bound to log model evidence $\mathcal{L}(q)$ as terms in Equation (12) are updated by our algorithm. The monotonic convergence of $\mathcal{L}(q)$ confirms the correctness of the iterative algorithm and its implementation.

We trained a HME model with a gating network comprising one gate and two experts. The feature weights of the gate are shown in Figure 3 for the sparse HME as well as the non-sparse HME. Sparse HME pushed most features to zero, while allowing a few features to take on large values. In contrast, non-sparse HME allowed several small non-zero feature weights. This result confirmed that indeed sparsity in the gate was obtained.

6. CONCLUSION

We presented the prior and a variational inference algorithm for the HME model that results in sparse gates and experts. Sparse HME was successfully deployed to build non-intrusive speech quality assessment algorithm. While sparse HME matched the performance of non-spars HME, the gate mechanism used fewer number of features.

7. REFERENCES

[1] M.I. Jordan and R.A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, vol. 6, no. 2, pp. 181–214, 1994.

[2] L. Breiman, *Classification and regression trees*, Chapman & Hall/CRC, 1984.

[3] C.M. Bishop and M. Svensén, "Bayesian hierarchical mixtures of experts," in *Proc. Nineteenth Conference on Uncertainty in Artificial intell.*, 2003, pp. 57–64.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[5] D.J.C. MacKay, "A practical Bayesian framework for backpropagation networks," *Neural Computation*, vol. 4, no. 3, pp. 448–472, 1992.

[6] R.M. Neal, *Bayesian Learning for Neural Networks*, Springer Verlag, 1996.

[7] "Single-ended method for objective speech quality assessment in narrow-band telephony applications," 2004, ITU, Geneva, Switzerland, 2004, ITU-T Rec. P.563.

[8] 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 *QoMEX*, 2010.

[9] "ITU-T coded-speech database," ITU, Geneva, Switzerland, 1998, ITU-T Rec. P.Suppl. 23.