A Bayesian hierarchical mixture of experts approach to estimate speech quality

A BAYESIAN HIERARCHICAL MIXTURE OF EXPERTS APPROACH TO ESTIMATE SPEECH QUALITY

S. Iman Mossavat, Oliver Amft, Bert de Vries *

Signal Processing Systems
Department of Electrical Engineering
Eindhoven University of Technology
The Netherlands
{i.mossavat, o.amft, b.de.vries}@tue.nl

Petko N. Petkov, W. Bastiaan Kleijn

Sound and Image Processing Lab
Department of Electrical Engineering
KTH-Royal Institute of Technology
Sweden
{petko.petkov, bastiaan.kleijn}@ee.kth.se

ABSTRACT

This paper demonstrates the potential of theoretically motivated learning methods in solving the problem of non-intrusive quality estimation for which the state-of-the-art is represented by ITU-T P.563 standard.

To construct our estimator, we adopt the speech features from P.563, while we use a different mapping of features to form quality estimates. In contrast to P.563 which assumes distortion-classes to divide the feature space, our approach divides the feature space based on a clustering which is learned from the data using Bayesian inference. Despite using weaker modeling assumptions, we are still able to achieve comparable accuracy on predicting mean-opinion-scores with P.563. Our work suggests Bayesian model-evidence as an alternative metric to correlation-coefficient for determining the necessary number of experts for modeling the data.

Index Terms— Speech, objective quality assessment, P.563, single-ended, non-intrusive, output-based, variational, Bayesian, mixture of experts

1. INTRODUCTION

Assessing the perceived quality of speech in communication channels is necessary to maintain the quality of service (QoS) and optimal usage of valuable network resources. While subjective quality assessment of speech is the most reliable method, its applicability is seriously constrained by time and cost, making the objective assessment of quality attractive. Objective speech quality assessment methods fall into two broad categories: *intrusive* such as , PSQM [1], MNB [2], and PESQ [3] and *non-intrusive* (a.k.a. single-ended or output-based) such as [4], [5], [6], [7] and [8]. While algorithms in the former category use a clean version of the speech signal to estimate the perceived quality of speech, the algorithms in the later category rely solely on the distorted speech signal. Intrusive methods typically predict the speech quality more accurately than their non-intrusive counterparts. Thus, intrusive methods are more suitable in the development stage of communication network equipment or algorithms where a clean copy of speech signal is readily available at no cost. Non-intrusive methods are more suitable in the deployment stage such as assessing the quality of live calls, where a clean version of the speech signal is not available (or costly to provide).

1.1. P.563 standard for non-intrusive speech quality assessment

The current ITU-T recommendation P.563 for non-intrusive speech quality assessment achieves high predictive accuracy [5]. P.563 design is based on the assumption that when different types of degradation occur simultaneously, humans focus on the dominant type. Distortions are categorized into six classes, *ordered* according to their annoyance: high level background noise (most annoying), signal interruptions, signal correlated noise, speech robotization, common unnaturalness (male talker), and common unnaturalness (female talker). The classification is implemented using a hierarchy of non-probabilistic (hard) binary classifiers. The binary classifiers compare *distortion-specific* features against corresponding thresholds. Starting from the binary classifier at the top of the hierarchy, a binary decision is made whether to classify the signal as the high-level background noise or proceed to the next classifier. The classification process continues down the hierarchy, until at some point the speech signal meets a classifier criteria. If the speech signal reaches the bottom of the hierarchy, it is classified as the common unnaturalness (female talker). In each distortion class, 12 features are linearly combined to form the intermediate quality estimation. After the intermediate quality estimation is calculated according to the distortion class, it is combined with additional features to form the final prediction [9] [5].

The design of P.563 is intuitive and interpretable. The

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

physical meanings of features play a significant role in how they appear in the classifiers, in the local linear predictors, or at the end of the mapping. The high accuracy of P.563 shows the classification is effective; However, the question remains whether the classification results in optimal prediction accuracy. It is also an interesting question to consider whether alternative ways of classification or clustering exist and what are their properties and implications. Furthermore, the entire classification decision in P.563 relies only on eight distortion-specific features out of the total of 43 features. Naturally, one may ask whether information from other features can also be used to divide the input space. Finally, P.563 divides the space into *hard* classes where each signal belongs to only one class. Thus at the class decision boundaries of P.563, abrupt changes are possible [10].

We would like to remind that according to Figure 9 of [5], the distribution of stimuli in the data-sets used to train and validate the performance of P.563 is not balanced, i.e, approximately 76% of the stimuli fall into two distortion-classes, and the share of the remaining four distortion-classes varies between 2.4% to 9% of the total number of stimuli.

1.2. Our Contribution

In this work, we demonstrate the applicability of the Bayesian framework to non-intrusive speech quality estimation. We keep the idea of dividing the higher dimensional feature space of P.563 into subregions. However, instead of accomplishing this by human intervention, we *learn* the sub-regions as well as the mapping in each sub-region. Using a statistical model, the hierarchical mixture of experts (HME) [11], we train a family of mappings that divide the input space in a *probabilistic* manner, i.e., our mappings allow the feature vector corresponding to a speech signal to lie at the same time in multiple sub-regions. Thus the transition from one region to another is gradual. An accurate approximate inference for Bayesian HME, referred to as variational Bayesian HME (VB-HME), forms the core of our modeling technique [12]. Using Bayesian methods, we are able to determine the number of sub-regions that are needed to model the data using HME, instead of presupposing it in advance as in P.563.

Finding the number of sub-regions in the data is a model selection problem. The correlation-coefficient is one of the most popular measures for model selection in non-Bayesian statistical methods. Our experiments show correlation-coefficient cannot provide enough evidence for selecting the number of sub-regions. In contrast, the Bayesian figure-of merit favors the simplest possible model with two sub-regions based on our data, which is in agreement with how P.563 classifier splits our data, i.e., ITU-T Supplement 23. While P.563 uses extensive knowledge from human experts, our Bayesian method relies on minimal assumptions.

The paper is organized as follows: in Section 2 we review the basic concepts of the Bayesian learning and define model-evidence, which serves as metric for Bayesian model selection. In Section 3 we present the HME model. We present the evaluation procedure in Section 4. In Section 5 we show our simulation results. We demonstrate how Bayesian model selection determines the number of experts in the ITU-T Supplement 23 database. Furthermore, we compare the predictive accuracy of HME with P.563 [9] and Bayesian MARS [13] methods. Finally Section 6 provides concluding remarks.

2. BAYESIAN LEARNING AND MODEL SELECTION

Given the feature vector $\boldsymbol{\psi}$ extracted from a speech signal, we wish to predict its quality y using the model $\mathcal{M}$ with parameters $\boldsymbol{\theta}$. We describe how to learn the parameters $\boldsymbol{\theta}$ and how to perform prediction based on training data-set $\mathcal{D}$ in the Bayesian framework, where $\mathcal{D} = \{(\boldsymbol{\psi}_i, y_i), i = 1, \cdots, N\}$, N is the number of data-points and y_i is the mean opinion score (MOS) obtained from subjective tests.

The first step in specifying a model in the Bayesian framework is to specify a *prior* over $p(\boldsymbol{\theta}|\mathcal{M})$ that encodes our knowledge about the parameters before seeing the data $\mathcal{D}$. The principle of maximum entropy is used in specifying priors for VB-HME to create non-informative priors [12]. This amounts to weak modeling assumptions, our fundamental modeling criteria. The next step is to define the *likelihood function*, $p(y|\boldsymbol{\psi}, \boldsymbol{\theta}, \mathcal{M})$, which specifies a distribution on the output y conditioned on the speech feature vector $\boldsymbol{\psi}$ and the parameter vector $\boldsymbol{\theta}$ of the model $\mathcal{M}$.

For a given model $\mathcal{M}$, the Bayes' formula is our main tool to update our prior knowledge about the parameters $\boldsymbol{\theta}$ in the light of the training data $\mathcal{D}$,

$$p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) = \frac{p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})}, \tag{1}$$

where $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})$ is called the *posterior* and represents the updated state of knowledge after observing $\mathcal{D}$ and $p(\mathcal{D}|\mathcal{M})$ is called the model-evidence and is defined shortly in Equation (2).

2.1. Model-Evidence

The *model-evidence* $p(\mathcal{D}|\mathcal{M})$ is computed by marginalizing the likelihood as follows

$$p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})\mathrm{d}\boldsymbol{\theta}. \tag{2}$$

Model-evidence does not play a critical role when performing predictions. Yet, if we want to perform Bayesian model selection the evidence serves as the metric that shows how well the model explains the data.

2.2. Bayesian Prediction

The Bayesian approach is fundamentally different from point estimate approaches, such as maximum-likelihood (ML) or

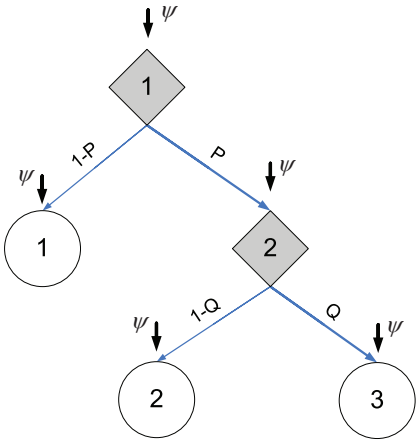


Fig. 1. A tree that represents a HME model with $C = 3$ experts and $M = 2$ gates. The gate are binary classifiers. In this illustration, gates one and two select the output branch on their right by probability P and Q respectively. The mixing coefficients at the leaves are computed by the product of the probabilities of the links on the path that connect the root to them.

the maximum-a-posteriori (MAP), in the sense that it represents our state of knowledge about the parameters $\boldsymbol{\theta}$ using the conditional *posterior* distribution $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})$ rather than a point estimate such as $\boldsymbol{\theta}_{\mathrm{ML}}$ or $\boldsymbol{\theta}_{\mathrm{MAP}}$. To perform predictions, we compute $p(y|\boldsymbol{\psi}, \mathcal{D}, \mathcal{M})$

$$p(y|\boldsymbol{\psi}, \mathcal{D}, \mathcal{M}) = \int p(y|\boldsymbol{\psi}, \boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) \mathrm{d}\boldsymbol{\theta}, \tag{3}$$

where the parameter is integrated over the posterior distribution. For a point-estimate, such as to compute the correlation-coefficient, we can use the mean or the mode of the $p(y|\boldsymbol{\psi}, \mathcal{D}, \mathcal{M})$ as our estimate. If we pick the mean as the estimate, the variance of the distribution can be used to measure our prediction confidence.

The integrations in Equations (3) and (2) are computationally challenging for high-dimensional feature spaces. Unlike many Bayesian methods that rely on demanding sampling methods for inference, an efficient and accurate method exists for approximating the posterior of VB-HME, as well as computing a tight lower bound on the log-model-evidence [12].

3. HIERARCHICAL MIXTURE OF EXPERTS

The HME describes a mixture distribution over the target variable y (the speech quality in our case) conditioned on the input feature vector $\boldsymbol{\psi}$. Figure 1 illustrates a hierarchical tree structures that is used to represent a typical HME model. The internal nodes of the tree are binary probabilistic classifiers called *gates* and the leaves are mixture components called *experts*. In Figure 1 there are two gates and three experts at the leaves. For each speech feature vector $\boldsymbol{\psi}$, the gates specify probability distributions on their output links. For example, gate one in Figure 1 chooses the right link with probability P, and the left link with probability $1 - P$. At each leaf, the mixing coefficient for the corresponding expert is computed by multiplying the link probabilities on the path that connects the root to the leaf. In Figure 1, for example, the mixing coefficients for experts one, two and three are $1 - P$, $P(1 - Q)$, and PQ respectively. Finally, each expert represents a probability distribution on y conditioned on $\boldsymbol{\psi}$. Thus, we have defined the following mixture distribution:

$$p(y|\boldsymbol{\psi}, \boldsymbol{\theta}_g, \{\boldsymbol{\theta}_k\}_{k=1}^{C}) = \sum_{k=1}^{C} g_k(\boldsymbol{\psi}|\boldsymbol{\theta}_g) p(y|\boldsymbol{\psi}, \boldsymbol{\theta}_k), \tag{4}$$

where C is the number of components or experts. The mixture parameters are divided into $\boldsymbol{\theta}_g$ that specifies the mixing coefficients $g_k(\boldsymbol{\psi}|\boldsymbol{\theta})$ and $\boldsymbol{\theta}_k$ for $k = 1, \cdots, C$ that specify the mixtures components (experts) $p(y|\boldsymbol{\psi}, \boldsymbol{\theta}_k)$. Note that both the mixing coefficients and the mixture components are conditioned on the speech features $\boldsymbol{\psi}$ meaning that for different regions in the space of $\boldsymbol{\psi}$, components have different weights in the mixture.

We suppose the expert distributions are Gaussian with mean $\boldsymbol{\omega}_k^{\mathrm{T}} \boldsymbol{\psi}$ and precision (inverse variance) τ_k for $k = 1, \cdots, C$, where vectors of feature weights $\boldsymbol{\omega}_k$ and precisions τ_k are to be learned from the data. Note that unlike hard binary classifiers in P.563, the gates in HME are probabilistic, thus , it is possible for speech samples to lie in several regions at the same time.

4. EVALUATION PROCEDURE

To evaluate the accuracy of our non-intrusive speech quality estimator, we use the Supplement 23 to the P series of ITU-T recommendations [14], henceforth referred to as P.Sup23. P.Sup23 constitutes of three experiments, where we use the absolute category rating (ACR) test results of experiments one and three. In the ACR method human listeners (subjects) are required to grade the speech samples on a discrete opinion scale: 'Excellent', 'Good','Fair','Poor','Bad'. For each sample, votes are averaged and referred to as mean opinion score (MOS) [15]. To compute the MOS for each speech file, the votes are averaged over 24 subjects. Our data-set constitutes of 1328 narrow-band speech signals, uniformly distributed over 50 distortion conditions. The speech features are extracted according to P.563 standard, and the target variable to predict is the corresponding MOS score. The data-set is divided into seven sub-databases that correspond to different labs and different experiments in P.Sup23 [14]. We removed 11 features out of the 43 P.563 features, since they did not vary much along our data-set. The last feature here is the bias, i.e, a constant number. Features are standardized.

Table 1. The distribution of the classification across speech databases of P.Sup23 [14] (1328 speech samples in ACR tests of experiments one and three)

P.563 Distortion Class	PSup.23
1 - High Level Background Noise	$< 1\%$
2 - Signal Interruptions	$< 1\%$
3 - Signal Correlated Noise	$< 1\%$
4 - Speech Robotization	6%
5 - Common Unnaturalness (Male)	45%
6 - Common Unnaturalness (Female)	47.5%

We train the VB-HME with different trees with $C = 2, \ldots, 5$ experts. For any number of experts given, we train on all the possible trees. To train the VB-HME, we use the deterministic-annealing procedure in [12] to initialize the parameters such that local optima are avoided. To compute the quality estimates we use the mean of $p(y|\boldsymbol{\psi}, \mathcal{D}, \mathcal{M})$, which yields higher prediction accuracy than the mode of the distribution in our experiments. We use a monotonic third-order polynomial as recommended in [5] to map the condition-averaged predictions to condition-averages MOS and compensate partially for factors, such as variation of scores in different languages. To perform cross-validation we use each database of the P.Sup23 as a test set and train on the 6 remaining databases.

5. SIMULATION RESULTS

In this section, we first consider the problem of determining the number of experts in VB-HME. We compare the behavior of two different metrics for model selection: model-evidence in Equation (2) and correlation-coefficient. Next, we compare our prediction results with P.563 and Bayesian MARS [13]. The major difference between the HME and the Bayesian MARS is that the HME is a parametric regression approach, whereas the Bayesian MARS is non-parametric. Thus, the Bayesian MARS requires far more memory for learning and prediction than the HME. Furthermore, the variational-Bayesian inference for HME converges much faster (in terms of time) compared to the demanding sampling method for the Bayesian MARS.

We compute the condition-averaged Pearson's correlation coefficient as in our previous work [13].

5.1. Explaining the data using VB-HME

Our simulations show that in the case of $C > 2$, the Bayesian inference shuts-down all the gates to **0** except the gate at the root of the tree. Once a gate is shut-down its two output branches are merged into one. Since only one expert at the root is non-zero, the effective number of sub-regions / experts is two for all trees in our experiments. In mixture-modeling, this effect is known as the component annihilation [16], which

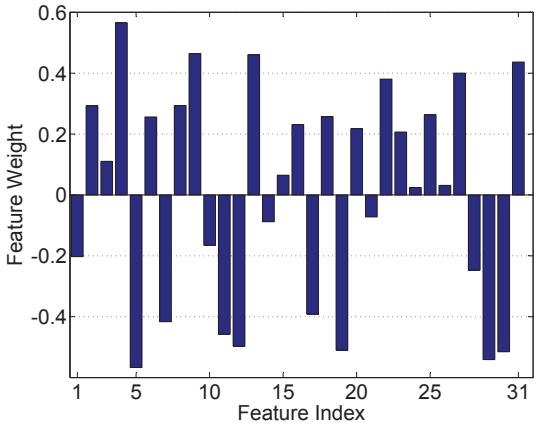


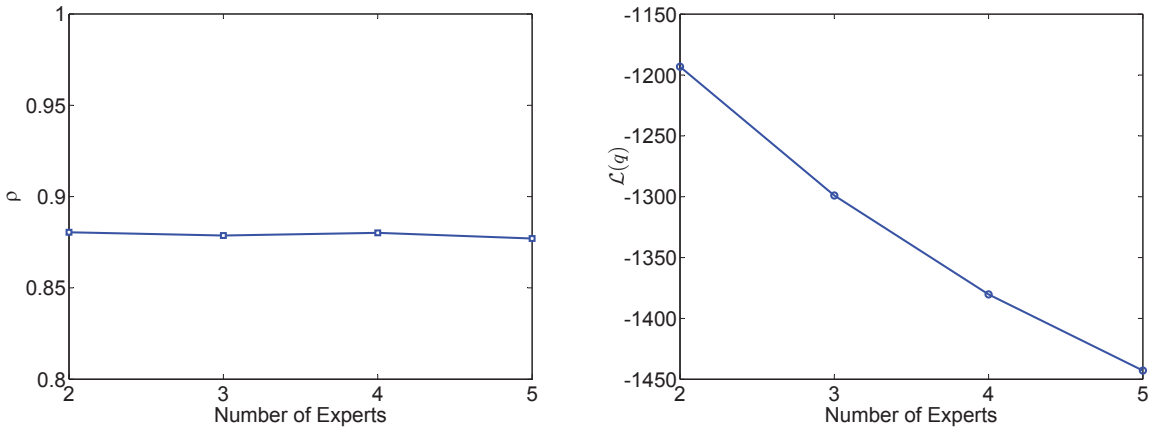
Fig. 2. Feature weights in the probabilistic binary classifier at the root of the tree.

reduces complex mixture-models to simpler ones. In contrast to the ML approach in which the extra model complexity leads to over-fitting, the Bayesian approach reduces the effective complexity of the model automatically. Table 1 shows that more than 92% of speech signals in our data-set lie only in two distortion-classes according to P.563 classifier: common unnaturalness for male and female talkers. Thus, the complexity of the solution based on VB-HME agrees with the assumptions incorporated in P.563.

By looking into the feature weights of the gate at the root as shown in Figure 2, we observe that the decision boundary of this gate involves almost all the features. P.563, however, uses one feature, averaged pitch, to classify the signals belonging to male and female talkers. Thus, the distortion classes of P.563 are different than the experts of the VB-HME. Our results in the the next section show that VB-HME slightly out-performs P.563 in predicting the speech quality.

As Figure 3a shows, the correlation-coefficient does not change as the number of experts changes. The reason is that all the models predict the data with equal accuracy. This happens since more complex models are simplified into simpler ones. In Figure 3b we illustrate the tight variational lower-bound $\mathcal{L}$ to the logarithm of model-evidence $\log p(\mathcal{D}|\mathcal{M})$. Higher values of $\mathcal{L}$, correspond to more plausible explanations of the data by the model. We observe that the Bayesian metric favors upon the simplest model with two experts.

The distinguishing aspect of model-evidence is that it offers a balanced metric between quality-of-fit and model complexity, whereas, correlation-coefficient does not incorporate model complexity explicitly. If two models predict the data with the same accuracy, the model-evidence favors simplicity while the correlation coefficient remains undecided. Thus the model-evidence is more compatible with the *Occamś razor* principle [17]: 'When two competing theories exist that make exactly the same predictions, the



(a) correlation-coefficient

(b) Lower-bound for log-model-evidence

Fig. 3. (a): The Pearson correlation-coefficient computed using cross-validation for VB-HME models with $C = 2, \ldots, 5$ experts. (b) Lower-bound for log-model-evidence, $\mathcal{L}(q)$, for VB-HME models with different number of experts. Symbols q and $\mathcal{L}(q)$ denote variational posterior approximation and lower-bound of log-model-evidence respectively, as defined in [12].

Table 2. Cross-validation results: comparison of predictive accuracy of Bayesian MARS [13] with P.563 features, P.563 [9], and VB-HME ($C = 2$) methods on 7 P.Sup23 [14] sub-databases (1328 speech files).

database	VB-HME	B-MARS	P.563
BNR-X1 (Canada)	0.949	0.935	0.911
BNR-X3	0.903	0.914	0.923
CNET-X1 (France)	0.870	0.861	0.798
CNET-X3	0.792	0.785	0.888
NTT-X1 (Japan)	0.898	0.933	0.867
NTT-X3	0.888	0.886	0.843
CSELT-X3 (Italy)	0.863	0.834	0.902
Mean	**0.881**	**0.878**	**0.876**

simpler one is the better.' It is important to note that these two metrics come from different theoretical frameworks and are not necessarily compatible, i.e., it is possible to draw different conclusions based on each of them. For example, once sufficient training data is available, changing priors does not influence the prediction results significantly (and consequently the correlation-coefficients remain unchanged), while priors influence model-evidence considerably.

5.2. Assessing Generalization Capability by Cross-validation

Following the discussion in Section 5.1 we choose the tree with $C = 2$ experts to compare VB-HME against P.563, and the Bayesian MARS algorithm by Petkov, Mossavat and Kleijn [13]. The Pearson correlation coefficient ρ for aforementioned algorithms over the 7 databases of experiment one and three of P.Sup23 are presented in Table 2. We observe that

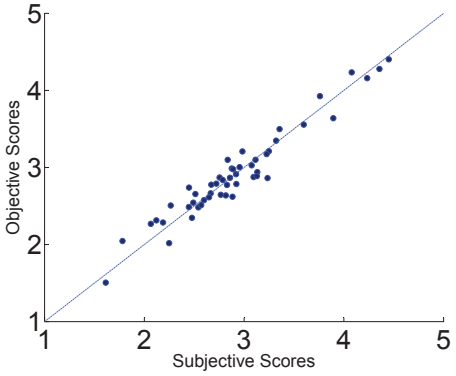


Fig. 4. Scatter plot of the condition-averaged mean-opinion-scores versus condition-averaged prediction results by the VB-HME. Each point corresponds to one of the distortion conditions in P.Sup23.

VB-HME slightly out-performs P.563. Figure 4 shows the scatter plot of condition-averaged mean-opinion-scores versus condition-averaged prediction results by VB-HME. Each point corresponds to one of the 50 distortion-conditions in P.Sup23. The prediction mean-squared-error increases as the points deviate from the line in the figure.

6. CONCLUSION AND FUTURE WORK

In this paper we demonstrate that while omitting the distortion-classes structure assumed by P.563 mapping, we achieve high accuracy in predicting the subjective mean-opinion-scores by using VB-HME. The structure of VB-HME resembles

P.563 in the sense that both methods work based on divide-and-conquer strategy, which splits the feature space into sub-spaces (sub-regions). In contrast to P.563, which *assumes* distortion-classes to form the sub-regions, VB-HME *learns* the sub-regions by training its gates, thus it is solving an unsupervised clustering problem based on weak assumptions (encoded in non-informative priors [12]). Thus, our work demonstrates how Bayesian learning can be used to explore the high-dimensional space of speech quality features based on statistical models.

We demonstrate the use of model-evidence as an alternative to correlation-coefficient for model selection. Model-evidence suggests that ITU-P.Sup23 data is best described by VB-HME using two experts, which is in agreement with P.563 distortion-classification results. In contrast to model-evidence, correlation-coefficient is relatively constant when different number of experts is used by VB-HME, i.e., correlation-coefficient cannot be used to determine the complexity of the model in this case. Note that, training an estimator is a fundamentally ill-posed problem and different assumptions can result in different results.

To look into the computational complexity of the HME, we note that for a given input $\boldsymbol{\psi}$, the HME computes the $p(y|\boldsymbol{\psi}, \mathcal{D}, \mathcal{M})$ over a grid of points in the range $[1-5]$, where for each point y_i in the grid, we have to compute predictions for two Bayesian linear regressors (for two experts), and one logistic function computation and one squaring-operation (for the gate).

More than 92% of speech samples in our data-set corresponds to only two out of six distortion-classes of P.563, i.e., speech samples in the top four distortion-classes of P.563 are relatively scarce. Thus, we need to test HME performance with larger data-sets in order to generalize our results. The results hold for the restricted case of conditions present in Supplement 23.

7. REFERENCES

[1] J. G. Beerends and J. A. Stemerdink, "A perceptual speech-quality measure based on a psychoacoustic sound representation," *J. of the Audio Engineering Society*, vol. 42, no. 3, pp. 115–123, 1994.

[2] S. Voran, "Objective Estimation of Perceived Speech Quality Part I: Development of the Measuring Normalizing Block Technique," *IEEE Trans. on Speech and Audio Proc.*, vol. 7, pp. 371–382, 1999.

[3] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual Evaluation of Speech Quality (PESQ): The New ITU Standard for End-to-End Speech Quality Assessment Part II-Psychoacoustic Model," *J. of the Audio Engineering Society*, vol. 50, no. 10, pp. 765–778, 2002.

[4] D. S. Kim, "ANIQUE: An Auditory Model for Single-ended Speech Quality Estimation," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5 Part 2, pp. 821–831, 2005.

[5] L. Malfait, J. Berger, and M. Kastner, "P. 563-the ITU-T Standard for Single-ended Speech Quality Assessment," in *IEEE Trans. on Audio, Speech and Language Proc.*, 2006, vol. 14, pp. 1924–1934.

[6] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-Complexity, Non-intrusive Speech Quality Assessment," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 14, no. 6, pp. 1948–1956, 2006.

[7] D. Picovici and A. E. Mahdi, "Output-based Objective Speech Quality Measure Using Self-organizing Map," in *IEEE Proceedings of ICASSP*, 2003, vol. 1, pp. 476–479.

[8] T. H. Falk and W. Chan, "Nonintrusive Speech Quality Estimation Using Gaussian Mixture Models," *IEEE Signal Processing Letters*, vol. 13, no. 2, pp. 108, 2006.

[9] ITU-T Recommendation P.563, "Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications," 2004.

[10] A. Ekman and W. B. Kleijn, "Improving Quality Prediction Accuracy of P.563 for Noise Suppression," in *Proc. Intl. Workshop Acoustic Echo and Noise Control*, 2008.

[11] 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.

[12] C. M. Bishop and M. Svensén, "Bayesian Hierarchical Mixtures of Experts," in *Proceedings Nineteenth Conference on Uncertainty in Artificial Intelligence*, 2003, pp. 57–64.

[13] 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 Language Processing*, Brighton, Sept. 2009, pp. 2875–2878.

[14] ITU-T, "Supplement 23 to the ITU-T P-series Recommendations: ITU-T Coded-speech Database," .

[15] ITU-T, "Mean Opinion Score (MOS) Terminology, ITU-T Recommendation P.800.1," .

[16] M.A.T. Figueiredo and A.K. Jain, "Unsupervised Learning of Finite Mixture Models," *IEEE Trans. on pattern analysis and machine intelligence*, pp. 381–396, 2002.

[17] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge Univ Pr, 2003.