EVALUATING INSTRUMENTAL MEASURES OF SPEECH QUALITY USING BAYESIAN MODEL SELECTION: CORRELATIONS CAN BE MISLEADING!

Antonio Kolossa, Johannes Abel, and Tim Fingscheidt

Institute for Communications Technology, Technische Universität Braunschweig, Germany

{kolossa, abel, fingscheidt}@ifn.ing.tu-bs.de

ABSTRACT

Choosing among competing models of collected data is crucial for all sciences. In the last decade there has been an increasing tendency to use Bayesian methods throughout many fields. When assessing the performance of instrumental measures of speech quality, classical measures such as correlation coefficients are still used. While these methods have their merits, they discard information about the data distribution, such as variability. They are useful as absolute measures of fit, but often not suitable for comparing different models. This paper uses Bayesian model selection, which does not suffer from these shortcomings, as it takes all information about the distribution of data into account and yields easily interpretable model probabilities. Two instrumental measures of speech quality are evaluated using data obtained in an absolute category rating (ACR) test. The results are compared and discussed. Bayesian methods prove superior for comparing instrumental measures, especially when the correlation of both measures is either poor or nearly identical. The proposed estimation procedure is highly recommended in selection phases for standardization bodies such as ITU-T, ETSI, 3GPP.

Index Terms: Bayesian model selection, instrumental speech quality measures

1. INTRODUCTION

Speech quality estimation via instrumental measures has been subject of research for many years. The motivation is the replacement of subjective listening tests, thus saving time and financial resources. Most commonly, the employed instrumental measures are intrusive, i.e., both, the reference speech signal (REF) as well as the coded (processed) speech signal (PROC) are needed for the quality prediction. This approach is also called full-reference signal-based model. Well-known examples are the wideband perceptual evaluation of speech quality (WB-PESQ) [1] algorithm and its successor perceptual objective listening quality assessment (POLQA) [2]. These two measures aim at modeling a subjective absolute category rating (ACR) listening test [3, Annex B].

Simplified, these algorithms predict speech quality in three steps. First, REF and PROC input signals are preprocessed in the frequency domain according to an approximation of human perception, which, e.g., considers non-linear loudness perception. A time (frame) and frequency (band) matrix $\Delta\mathbf{P} = \mathbf{P}_{\text{REF}} - \mathbf{P}_{\text{PROC}}$ is calculated, representing the perceptual difference due to transcoding. In the second step, $\Delta\mathbf{P}$ is integrated over time and frequency, leading to a single scalar value per utterance, which is then mapped in the third step via a pre-trained linear regression to the mean opinion score listening quality objective (MOS-LQO). The linear regression coefficients are found using training material consisting of speech signal pairs for REF and PROC in combination with the according mean opinion score listening quality subjective (MOS-LQS) values, obtained beforehand in subjective listening tests.

The absolute performance of an instrumental measure is often given as Pearson's correlation coefficient [4], which is a measure of the similarity between MOS-LQO and MOS-LQS values, thus indicating how well the instrumental measures approximate the subjective listening test. However, when comparing measures [5] or different tunings from the same measure during development [6], a final conclusion based on correlation coefficients might not be possible. This is the case, if the compared correlation factors are very small, or their values are very close to each other [7]. Even if the coefficients diverge, interpretation of the significance of the difference between the models is not intuitive.

This paper introduces Bayesian model selection (BMS) to compare the predictions from different instrumental measures of speech quality directly with each other based on easily interpretable posterior model probabilities. An existing subjective listening test was used in order to evaluate MOS-LQO values obtained from both WB-PESQ and POLQA using BMS, which offers a framework for comparing models of measured data while taking inter- and intra-person variability into account [8, 9]. These methods are applied in statistics [10–12] and are proven as useful tools for model selection in many applications [13], among them signal processing [14], machine learning [15], natural language processing [16], neuroimaging [17, 18], social sciences [19], and biology [20, 21].

In this paper, correlation coefficients and posterior model probabilities based on Bayesian estimation are assessed as measures for model selection. The information contained in correlation coefficients and posterior model probabilities as well as the different views on collected data inherent in typical test setups is discussed.

The paper is structured as follows: The next section provides an introduction to Bayesian model selection and linear hierarchical models. After presenting the experimental setup in Section 3 and results in Section 4, conclusions are finally drawn in Section 5.

2. BAYESIAN MODEL SELECTION USING POSTERIOR MODEL PROBABILITIES

To compare different instrumental measures which yield MOS-LQO values with the obtained MOS-LQS values a general linear hierarchical model is used in the form of parametric empirical Bayesian (PEB) schemes[1]. In this section, posterior model probabilities used for Bayesian model selection (BMS) are derived. A detailed specification of the employed hierarchical model is given later on in Section 3.3.

[1]The implementation of the PEB schemes is freely available in the statistical parametric mapping (SPM) software (`spm_PEB.m`) [9, 22].

$\mathrm{P}(m\|\mathbf{y})$	0.50 - 0.75	0.75 - 0.95	0.95 - 0.99	>0.99
Significance	weak	positive	strong	very strong

Table 1. Significance of posterior model probabilities [8, 23].

Empirical Bayes models equip a general linear model with further hierarchical levels that place constraints on the parameters of the lower levels. An expectation maximization algorithm estimates all unknown model parameters to calculate the variational free energy F which consists of an accuracy and a complexity term [22–24]. F_m is a lower bound on the marginal log-likelihood or log-evidence $\ln(\mathrm{p}(\mathbf{y}|m))$, with $\mathrm{p}(\mathbf{y}|m)$ being the likelihood of the data $\mathbf{y}$ (MOS-LQS values) given the model $m \in \mathcal{M} = \{\text{WB-PESQ, POLQA}\}$ to compute the MOS-LQO. The evidence is approximated using this lower bound by $\mathrm{p}(\mathbf{y}|m) \approx e^{F_m}$ [13].

Given the model likelihoods $\mathrm{p}(\mathbf{y}|m)$, the models $m \in \mathcal{M}$ can be compared via posterior model probabilities $\mathrm{P}(m|\mathbf{y})$ which are calculated following Bayes' theorem

$$\mathrm{P}(m|\mathbf{y}) = \frac{\mathrm{p}(\mathbf{y}|m)\mathrm{P}(m)}{\sum\limits_{\mu\in\mathcal{M}} \mathrm{p}(\mathbf{y}|\mu)\mathrm{P}(\mu)}, \tag{1}$$

with $\mathrm{P}(m)$ being an *a priori* model probability [13]. Assuming that all models are equally probable *a priori*, (1) is simplified to

$$\mathrm{P}(m|\mathbf{y}) = \frac{\mathrm{p}(\mathbf{y}|m)}{\sum\limits_{\mu\in\mathcal{M}} \mathrm{p}(\mathbf{y}|\mu)} \approx \frac{e^{F_m}}{\sum\limits_{\mu\in\mathcal{M}} e^{F_\mu}}. \tag{2}$$

Table 1 gives an overview on how to interpret the significance of posterior model probabilities. Values larger than 0.5 reflect evidence in favor of model m, with values larger than 0.95 being considered as 'strong', and 0.99 as 'very strong' evidence [8, 23].

3. EXPERIMENTAL SETUP

In this section a brief description of an example ACR listening test setup is given, followed by the evaluation strategy. Finally, we introduce the linear hierarchical model and provide a detailed specification of the required design matrices for Bayesian model selection.

3.1. An Example ACR Listening Test Setup

A typical ACR listening test was used as basis for the experiments. The test will now be sketched, a detailed description can be found in [25]. Several algorithms for artificial speech bandwidth extension (ABE) served as test objects, but keep in mind: Our particular focus in this paper is not to find out, which of the these ABE algorithms performs best, but which of the models $m \in \mathcal{M} = \{\text{WB-PESQ, POLQA}\}$ provides the best predictions for the MOS-LQS.

Speech data for the listening test was taken from the German part of the NTT-AT database for telephonometry [26]. Two female and two male speakers were selected, each speaker providing four sentences. The subjects were asked to give an absolute rating scale in MOS-LQS from 1 (bad) to 5 (excellent) for every file. In the listening test $C = 16$ conditions $c \in \mathcal{C}_{\text{ALL}} = \{1, \ldots, C\}$ were presented to $L = 24$ subjects: one coded narrowband (NB) ($\mathcal{C}_{\text{AMR-NB}} = \{1\}$), six ABE ($\mathcal{C}_{\text{ABE}} = \{2, \ldots, 7\}$), three coded wideband (WB) ($\mathcal{C}_{\text{AMR-WB}} = \{8, 9, 10\}$), and six WB modulated noise reference unit (MNRU) [31] conditions ($\mathcal{C}_{\text{MNRU}} = \{11, \ldots, 16\}$).

One out of the four utterances per speaker was used during a preliminary familiarization phase, leaving a total of $N = 12$ sentences per condition. In total $C \cdot N = 192$ files were evaluated in the ACR test.

Speech data sampled at 16 kHz was selected for all conditions. For the NB condition, the speech data was decimated using a FLAT1 filter with 3.6 kHz cutoff frequency [27] and the adaptive multirate NB (AMR-NB) speech codec [29] was applied at a bitrate of 12.2 kbps. This NB condition also serves as input to the six different ABE algorithms. Coded WB conditions were obtained by transcoding the input speech via the adaptive multirate WB (AMR-WB) speech codec [30] at bitrates of 8.85 kbps, 12.65 kbps, and 23.85 kbps. The remaining six WB conditions were generated by using the MNRU with speech to modulated noise power ratios of ∞ dB (clean), 45 dB, 35 dB, 25 dB, 15 dB, and 5 dB. Detailed test results are discussed in [25] but are not of particular interest here.

3.2. Evaluation

In order to compare correlation coefficients to the outcome of Bayesian model selection, a file-based evaluation has to be done. To obtain subjective votes on file-basis, it is necessary to average over the different subjects $\ell \in \mathcal{L} = \{1, \ldots, L\}$ for each condition c and sentence n following

$$y_{c,n} = \frac{1}{|\mathcal{L}|} \sum_{\ell\in\mathcal{L}} y_{c,n,\ell}, \quad n \in \mathcal{N} = \{1, ..., N\},$$

which eliminates inter-person variance, with $|\mathcal{L}| = L$ being the number of subjects. The MOS-LQO values $x_{c,n}$ which correspond to the resulting $y_{c,n}$ are calculated using WB-PESQ and POLQA. While the different conditions constitute the PROC cases, the respective WB MNRU with ∞ dB modulated noise power ratio is the REF case. POLQA operates in superwideband (SWB) mode, therefore the speech files have been interpolated to 32 kHz sampling rate in advance.

For evaluation, the $C = 16$ conditions were clustered into the five group sets $\mathcal{C}_g$ with $g \in \mathcal{G} = \{\text{ALL,AMR-NB, ABE, AMR-WB, MNRU}\}$. The file-based correlation coefficient [4, 34]

$$r_g = \frac{\sum\limits_{c\in\mathcal{C}_g}\sum\limits_{n\in\mathcal{N}} \left(x_{c,n} - \overline{x}\right)\left(y_{c,n} - \overline{y}\right)}{\sqrt{\left(\sum\limits_{c\in\mathcal{C}_g}\sum\limits_{n\in\mathcal{N}} \left(x_{c,n} - \overline{x}\right)^2\right)\left(\sum\limits_{c\in\mathcal{C}_g}\sum\limits_{n\in\mathcal{N}} \left(y_{c,n} - \overline{y}\right)^2\right)}}, \tag{3}$$

with $\overline{x}$ and $\overline{y}$ denoting the global mean of all MOS-LQO and MOS-LQS values, respectively, was calculated for every group set $\mathcal{C}_g$. The root-mean-square error (RMSE) between $x_{c,n}$ and $y_{c,n}$ is also calculated for each group set $\mathcal{C}_g$.

3.3. Hierarchical Linear Model

In the following, we give an introduction to the linear hierarchical model as used here for Bayesian model selection and a detailed specification of the design matrices and data vectors, which are required to calculate the model evidence F_m in (2). The different models (WB-PESQ, POLQA) generate the model-specific MOS-LQO values as regressors $x_{n,\ell}$, with the subscript ℓ denoting the individual test subjects and n being the consecutive index of the sentences which were actually presented to a subject within one test condition c. The condition index c is omitted in the general description of the hierarchical model for better readability: it actually applies to all variables until eq. (6). The first level of the hierarchical model assumes the MOS-LQS values $y_{n,\ell}$ to be *directly proportional* to the MOS-LQO values $x_{n,\ell}$ with the unknown parameter $\theta_\ell^{(1)}$ and error $\epsilon_{n,\ell}^{(1)}$

$$y_{n,\ell} = x_{n,\ell}\theta_\ell^{(1)} + \epsilon_{n,\ell}^{(1)}, \tag{4}$$

with superscript $^{(j)}$ denoting the level $j \in \mathcal{J} = \{1, 2, 3\}$. Note that the first level (4) models no deterministic offset as both WB-PESQ and POLQA are specifically designed to directly substitute the obtained MOS-LQS. The second level models the *subject-specific* parameters $\theta_\ell^{(1)}$ as deviations from a group parameter $\theta^{(2)}$

$$\theta_\ell^{(1)} = \theta^{(2)} + \epsilon_\ell^{(2)}. \tag{5}$$

The third level specifies no prior knowledge for the group parameter $\theta^{(2)} = \epsilon^{(3)}$ [9]. In the following, non-bold letters refer to scalars, bold small letters to (column) vectors, and bold capital letters to matrices. In this notation the complete general hierarchical model is

$$\begin{aligned} \mathbf{y} &= \mathbf{X}^{(1)}\boldsymbol{\theta}^{(1)} + \boldsymbol{\epsilon}^{(1)} \\ \boldsymbol{\theta}^{(1)} &= \mathbf{x}^{(2)}\theta^{(2)} + \boldsymbol{\epsilon}^{(2)} \\ \theta^{(2)} &= \epsilon^{(3)}. \end{aligned} \tag{6}$$

For evaluation of the groups $g \in \mathcal{G}$, the respective sets of conditions $\mathcal{C}_g$, with $C_g = |\mathcal{C}_g|$ being the number of conditions in group g, yield the respective data vector $\mathbf{y}_g \in \mathbb{R}^{C_g NL}$ according to $\mathbf{y}_g = [\mathbf{y}_{g,\ell=1}^T, ..., \mathbf{y}_{g,\ell=L}^T]^T$, with $\mathbf{y}_{g,\ell} = [y_{c_{g,\min},n=1,\ell}, ..., y_{c_{g,\min},n=N,\ell}, ..., y_{c_{g,\max},n=N,\ell}]^T \in \mathbb{R}^{C_g N}$ and $(\,)^T$ being the transpose. Here, $c_{g,\min}$ and $c_{g,\max}$ denote $\min(c)$ and $\max(c)$ with $c \in \mathcal{C}_g$, respectively. The first level design matrix $\mathbf{X}_g^{(1)} \in \mathbb{R}^{C_g NL \times L}$ is block-diagonal with L partitions $\mathbf{x}_{g,\ell} = [x_{c_{g,\min},n=1,\ell}, ..., x_{c_{g,\max},n=N,\ell}]^T \in \mathbb{R}^{C_g N}$ containing the MOS-LQO values of one subject ℓ.

The second level design matrix $\mathbf{X}^{(2)} = \mathbf{x}^{(2)} = \mathbf{1}_L$ is an all-one column vector of length L, expressing that the subject specific parameters $\theta_\ell^{(1)}$ are deviations of a single subject-independent group parameter $\theta^{(2)}$ (5). For well-fitted models, the group parameter $\theta^{(2)}$ should be close to one. Technically, the third level contains a design matrix which is set to the scalar value of zero $\mathbf{X}^{(3)} = x^{(3)} = 0$. Notice that the vector $\mathbf{y}$ and matrices $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$, and $\mathbf{X}^{(3)}$ form the only model-specific input to the PEB method [9, 22] for estimating the unknown parameters and calculating F_m for use in (2).

The parameter vector $\boldsymbol{\theta}^{(1)} = [\theta_{\ell=1}^{(1)}, ..., \theta_{\ell=L}^{(1)}]^T \in \mathbb{R}^L$ assembles the unknown level-one parameters $\theta_\ell^{(1)}$, which are the link between MOS-LQO and MOS-LQS (4). All parameters are treated as multivariate random Gaussian variables with posterior densities $\mathcal{N}(\boldsymbol{\theta}^{(j)}; \boldsymbol{\mu}_{\theta|y}^{(j)}, \boldsymbol{\Sigma}_{\theta|y}^{(j)})$. The conditional means of these densities $\boldsymbol{\mu}_{\theta|y}^{(j)}$ are used as point estimates of the parameters $\boldsymbol{\theta}^{(j)}$ [9]. All errors are assumed to be normally distributed $\boldsymbol{\epsilon}^{(j)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon^{(j)})$. The covariance is parameterized by the hyperparameters $\lambda^{(j)}$ following $\boldsymbol{\Sigma}_\epsilon^{(j)} = \lambda^{(j)}\mathbf{I}^{(j)}$, with covariance constraint $\mathbf{I}^{(j)}$ being an identity matrix with the same dimensions as the number of rows of the design matrix $\mathbf{X}^{(j)}$ of the corresponding level j. The unknown parameters $\boldsymbol{\theta}^{(j)}$ and hyperparameters $\lambda^{(j)}$ are estimated using an expectation maximization algorithm.

3.4. Parametric Empirical Bayes Computation

Given the data vector $\mathbf{y}$ and design matrices $\mathbf{X}^{(j)}$ for all levels $j \in \mathcal{J}$ along with the error covariance constraints $\mathbf{I}^{(j)}$ from Section 3.3, the parameter estimates $\boldsymbol{\mu}_{\theta|y}^{(j)}$, error covariance matrices $\boldsymbol{\Sigma}_\epsilon^{(j)}$, and model evidence F_m are calculated based on expectation maximization by the PEB scheme: $[\boldsymbol{\mu}_{\theta|y}^{(\mathcal{J})}, \boldsymbol{\Sigma}_\epsilon^{(\mathcal{J})}, F_m] = \mathtt{PEB}(\mathbf{y}, \mathbf{X}^{(\mathcal{J})}, \mathbf{I}^{(\mathcal{J})})$. The obtained F_m for all $m \in \mathcal{M}$ are used for calculating the posterior model probabilities $\mathrm{P}(m|\mathbf{y})$ after (2), which allows for direct Bayesian model selection following Table 1.

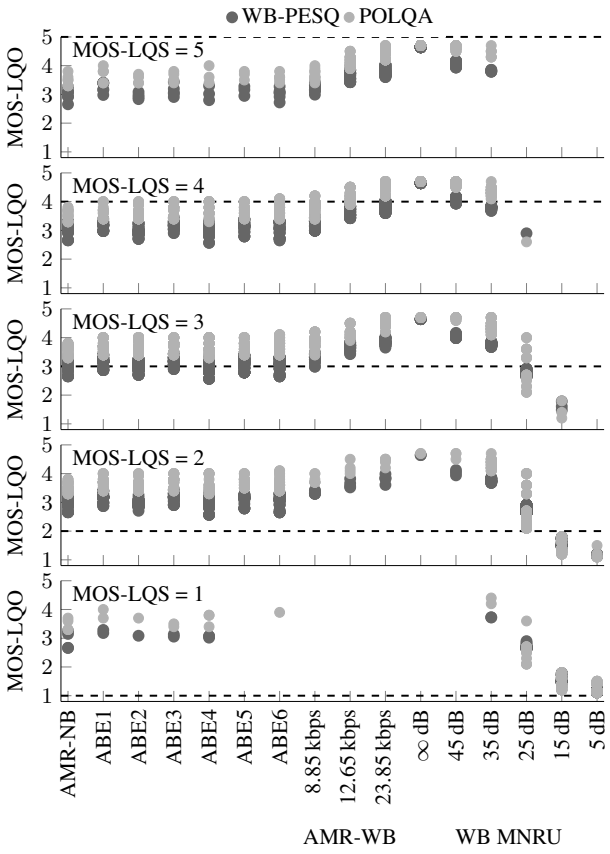

Fig. 1. Detailed scatter plot of WB-PESQ (●) and POLQA (●) sorted by obtained MOS-LQS from MOS-LQS = 5 (top panel) down to MOS-LQS = 1 (bottom panel).

4. EXPERIMENTAL RESULTS

Figure 1 shows a detailed scatter plot of MOS-LQO values obtained from WB-PESQ (●) and POLQA (●) over conditions, sorted by the obtained MOS-LQS values. Once a file gets different scores from different subjects, it consequently appears in more than one of the subplots. Especially for MOS-LQS values between 2 and 5, similarities in the plots are apparent that prove high inter-person variability over most of the conditions. Throughout all plots, MOS-LQO values from POLQA are higher than those from WB-PESQ.

In Figure 2 the file-based MOS-LQS values $y_{c,n}$ for $c \in \mathcal{C}_{\text{ALL}}$ are plotted over the corresponding MOS-LQO values $x_{c,n}$ from WB-PESQ (●) and POLQA (●). As in Figure 1 systematically higher scores are given by POLQA. Additionally, a strong linear dependency is suggested.

Table 2 shows posterior model probabilities $\mathrm{P}(m|\mathbf{y})$ after (2), correlation coefficients r_g after (3), second-level parameters $\theta^{(2)}$ (5), and RMSE values for WB-PESQ and POLQA for all groups $g \in \mathcal{G}$. High correlation coefficients of $r_{\text{ALL}} = 0.90$ for WB-PESQ and $r_{\text{ALL}} = 0.89$ for POLQA for the group set $\mathcal{C}_{\text{ALL}}$ confirm the overall impression from Figure 2. However, posterior model probabilities clearly state that WB-PESQ predicts MOS values *of this* subjective test better than POLQA, with $\mathrm{P}(m = \text{WB-PESQ}|\mathbf{y}) > 0.99$, i.e.,

	$m=$WB-PESQ				$m=$POLQA			
Condition group set $\mathcal{C}_g$	$\mathrm{P}(m\|\mathbf{y})$	r_g	$\theta^{(2)}$	RMSE	$\mathrm{P}(m\|\mathbf{y})$	r_g	$\theta^{(2)}$	RMSE
$\mathcal{C}_{\text{ALL}}$	>0.99	0.90	0.98	0.40	<0.01	0.89	0.86	0.68
$\mathcal{C}_{\text{AMR-NB}}$	0.51	0.31	0.95	0.31	0.49	0.24	0.82	0.68
$\mathcal{C}_{\text{AMR-WB}}$	>0.99	0.21	1.07	0.51	<0.01	0.08	0.93	0.54
$\mathcal{C}_{\text{ABE}}$	0.65	0.08	0.98	0.29	0.35	0.09	0.82	0.71
$\mathcal{C}_{\text{MNRU}}$	>0.99	0.96	0.92	0.45	<0.01	0.94	0.85	0.70

Table 2. Comparison of posterior model probabilities $\mathrm{P}(m|\mathbf{y})$ (2), correlation coefficients r_g (3), group parameters $\theta^{(2)}$ (6) of the hierarchical model, and root-mean-square error (RMSE) for WB-PESQ and POLQA. Values are shown for all condition group sets.

very strong statistical significance. Notice that due to the high correlation coefficients, both measures prove to represent human perception well in absolute terms.

POLQA tends to overestimate speech quality (as shown in Figures 1 and 2) which is also expressed in $\theta^{(2)} = 0.86 < 1$. WB-PESQ for this test proves to be more accurately mapped towards the MOS-LQS scale, with $\theta^{(2)} = 0.98 \approx 1$. This result is confirmed by a significantly lower RMSE for WB-PESQ (0.40) compared to POLQA (0.68).

Table 2 also shows the results for the other groups of conditions. With respect to coded WB group set $\mathcal{C}_{\text{AMR-WB}}$, Bayesian model selection (BMS) puts WB-PESQ in favor of POLQA. This is confirmed by the significant difference of the correlation coefficients of 0.21 for WB-PESQ and 0.08 for POLQA. However, both measures have low correlation coefficients and are therefore not able to accurately predict MOS-LQO values for these conditions. Interestingly, WB-PESQ slightly underestimates the MOS-LQS values in these conditions, as revealed by $\theta^{(2)} = 1.07 > 1$, while POLQA overestimates the MOS-LQS with $\theta^{(2)} = 0.93 < 1$. The size of underestimation by WB-PESQ and overestimation by POLQA is about equal, which is directly mirrored in the similar RMSEs of 0.51 and 0.54, respectively. Regarding the coded NB speech group set $\mathcal{C}_{\text{AMR-NB}}$, neither WB-PESQ nor POLQA (in SWB mode) have high correlation coefficients and BMS states about equal posterior model probability for WB-PESQ and POLQA. The differences of the parameters $\theta^{(2)}$ are again directly mirrored by the RMSE.

The six ABE conditions $\mathcal{C}_{\text{ABE}}$ are poorly represented by both instrumental measures. Correlation factors are $r_{\text{ABE}} = 0.08$ and $r_{\text{ABE}} = 0.09$ for WB-PESQ and POLQA, respectively. A posterior model probability of $\mathrm{P}(m = \text{WB-PESQ}|\mathbf{y}) = 0.65$ signifies only weak statistical evidence in favor of WB-PESQ. Significant differences between $\theta^{(2)}$ and the RMSE strengthen the overall impression of a close relationship between these two values.

For WB MNRU conditions $\mathcal{C}_{\text{MNRU}}$, WB-PESQ and POLQA show very high correlation factors of $r_{\text{MNRU}} = 0.96$ and $r_{\text{MNRU}} = 0.94$, respectively. *Even though these correlation coefficients are very similar, BMS clearly favors WB-PESQ with a posterior model probability of* $\mathrm{P}(m = \textit{WB-PESQ}|\mathbf{y}) > 0.99$*!* The parameter $\theta^{(2)}$ is closer to one and the RMSE is smaller for WB-PESQ than for POLQA, thus confirming the results obtained via posterior model probabilities.

5. CONCLUSIONS

This paper introduces Bayesian model selection (BMS) for the evaluation of instrumental measures of speech quality. Further, posterior model probabilities are compared to correlation coefficients with regard to their explanatory power and the relationship between Bayesian parameter estimates and the RMSE is investigated. As an example, WB-PESQ and POLQA serve as models for data obtained in an absolute category rating test. Although the results from BMS and correlation are qualitatively identical, there is a tremendous difference in statistical significance. Specifically, in conditions with high and nearly identical correlation coefficients, posterior model probabilities state very strong evidence in favor of WB-PESQ vs. POLQA. For some conditions, however, the correlation coefficients are small but different and BMS states only weak superiority of WB-PESQ. The RMSE confirms the results of the BMS and the differences in RMSE directly mirror the Bayesian parameter estimates.

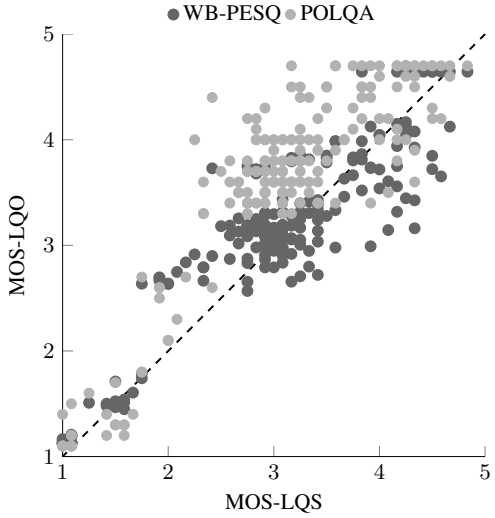

Fig. 2. Detailed scatter plot of WB-PESQ (●) and POLQA (●) over file-based mean MOS-LQS for group set $\mathcal{C}_{\text{ALL}}$.

As *absolute* measures of fit, a correlation coefficient and the RMSE serve their purpose very well by directly relating objective to subjective mean opinion scores. However, if speech codecs or instrumental measures of speech quality need to be *compared*, e.g. in a selection phase of a standardization body such as ITU-T, ETSI or 3GPP, the correlation fails, since it only takes *mean* scores into account and is therefore blind to intra- and inter-subject variability. By using Bayesian model selection, these variabilities are considered, thus leading to more meaningful and significant comparisons. Furthermore, Bayesian parameter estimates indicate systematic errors due to offsets which cannot be directly seen in the RMSE values.

6. REFERENCES

[1] "ITU-T Recommendation P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," ITU, Nov. 2007.

[2] "ITU-T Recommendation P.863, Perceptual Objective Listening Quality Assessment," ITU, Jan. 2011.

[3] "ITU-T Recommendation P.800, Methods for Subjective Determination of Transmission Quality," ITU, Aug. 1996.

[4] "ITU-T Recommendation P.862, Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs," ITU, Feb. 2001.

[5] S. Möller, E. Kelaidi, F. Köster, N. Côté, P. Bauer, T. Fingscheidt, T. Schlien, H. Pulakka, and P. Alku, "Speech Quality Prediction for Artificial Bandwidth Extension Algorithms," in *Proc. of INTERSPEECH 2013*, Lyon, France, Aug. 2013, pp. 3439–3443.

[6] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ) - A new Method for Speech Quality Assessment of Telephone Networks and Codecs," in *Proc. of ICASSP 2001*, Salt Lake City, Utah, USA, May 2001, IEEE, vol. 2, pp. 749–752.

[7] N. Côté, *Integral and Diagnostic Intrusive Prediction of Speech Quality.*, T-Labs Series in Telecommunication Services. Springer, 2011.

[8] R. E. Kass and A. E. Raftery, "Bayes Factors," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995.

[9] K. J. Friston, W. D. Penny, C. Phillips, S. J. Kiebel, G. Hinton, and J. Ashburner, "Classical and Bayesian Inference in Neuroimaging: Theory," *NeuroImage*, vol. 16, no. 2, pp. 465–483, 2002.

[10] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian Model Averaging: A Tutorial," *Statistical Science*, vol. 14, no. 4, pp. 382–401, 1999.

[11] M. A. Pitt and I. J. Myung, "When a Good Fit Can be Bad," *Trends in Cognitive Sciences*, vol. 6, no. 10, pp. 421–425, 2002.

[12] H. Hoijtink, I. Klugkist, and P. A. Boelen, *Bayesian Evaluation of Informative Hypotheses*, Springer, NY, 2008.

[13] W. D. Penny, K. E. Stephan, J. Daunizeau, M. J. Rosa, K. J. Friston, T. M. Schofield, and A. P. Leff, "Comparing Families of Dynamic Causal Models," *PLoS Computational Biology*, vol. 6, no. 3, pp. e1000709, 2010.

[14] W. D. Penny and S. J. Roberts, "Bayesian Multivariate Autoregressive Models with Structured Priors," *IEE Proc. - Vision, Image and Signal Processing*, vol. 149, no. 1, pp. 33–41, 2002.

[15] M. J. Beal and Z. Ghahramani, *The Variational Bayesian EM Algorithm for Incomplete Data: With Application to Scoring Graphical Model Structures. In: Bayesian Statistics 7 (Bernado J. M., Bayarri M. J., Berger J. O., Dawid A. P., Heckerman D., Smith A. F. M., West M., eds.)*, pp. 453–464, Oxford University Press, Oxford, UK, 2003.

[16] C. Kemp, A. Perfors, and J. B. Tenenbaum, "Learning Overhypotheses with Hierarchical Bayesian Models," *Developmental Science*, vol. 10, no. 3, pp. 307–321, 2007.

[17] M. W. Woolrich, "Bayesian Inference in fMRI," *NeuroImage*, vol. 62, no. 2, pp. 801–810, 2012.

[18] A. Kolossa, B. Kopp, and T. Fingscheidt, "A Computational Analysis of the Neural Bases of Bayesian Inference," *NeuroImage*, vol. 106, no. 21, pp. 222–237, 2015.

[19] A. E. Raftery, "Bayesian Model Selection in Social Research," *Sociological Methodology*, vol. 25, pp. 111–164, 1995.

[20] V. Vyshemirsky and M. A. Girolami, "Bayesian Ranking of Biochemical System Models," *Bioinformatics*, vol. 24, no. 6, pp. 833–839, 2008.

[21] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf, "Approximate Bayesian Computation Scheme for Parameter Inference and Model Selection in Dynamical Systems," *Journal of the Royal Society Interface*, vol. 6, no. 31, pp. 187–202, 2009.

[22] K. J. Friston, J. Mattout, N. Trujillo-Bareto, J. Ashburner, and W. D. Penny, "Variational Free Energy and the Laplace Approximation," *NeuroImage*, vol. 34, no. 1, pp. 220–234, 2007.

[23] W. D. Penny, K. E. Stephan, A. Mechelli, and K. J. Friston, "Comparing Dynamic Causal Models," *NeuroImage*, vol. 22, no. 3, pp. 1157–1172, 2004.

[24] W. D. Penny, "Comparing Dynamic Causal Models Using AIC, BIC and Free Energy," *NeuroImage*, vol. 59, no. 1, pp. 319–330, 2012.

[25] P. Bauer, J. Abel, and T. Fingscheidt, "HMM-Based Artificial Bandwidth Extension Supported by Neural Networks," in *Proc. of International Workshop on Acoustic Signal Enhancement*, Juan les Pins, France, Sept. 2014, pp. 1–5.

[26] "Multi-Lingual Speech Database for Telephonometry," NTT Advanced Technology Corporation (NTT-AT), 1994.

[27] "ITU-T Recommendation G.191, Software Tool Library 2009 User's Manual," ITU, Nov. 2009.

[28] "ITU-T Recommendation P.56, Objective Measurement of Active Speech Level," ITU, Dec. 2011.

[29] "Mandatory Speech Codec Speech Processing Functions: AMR Speech Codec; Transcoding Functions (3GPP TS 26.090, Rel. 6)," 3GPP; TSG SA, Dec. 2004.

[30] "Speech Codec Speech Processing Functions: AMR Wideband Speech Codec; Transcoding Functions (3GPP TS 26.190, Rel. 6)," 3GPP; TSG SA, Dec. 2004.

[31] "ITU-T Recommendation P.810, Modulated Noise Reference Unit (MNRU)," ITU, Feb. 1996.

[32] H. Pulakka and P. Alku, "Bandwidth Extension of Telephone Speech Using a Neural Network and a Filter Bank Implementation for Highband Mel Spectrum," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2170–2183, 2011.

[33] "ITU-T Recommendation P.341, Transmission Characteristics for Wideband Digital Loudspeaking and Hands-Free Telephony Terminals," ITU, Mar. 2011.

[34] "Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality Performance in the Presence of Background Noise Part 3: Background Noise Transmission - Objective Test Methods," ETSI, Nov. 2008.