FEATURE SET AUGMENTATION FOR ENHANCING THE PERFORMANCE OF A NON-INTRUSIVE QUALITY PREDICTOR

Petko N. Petkov[1]*, Hannes Helgason*[1]*, W. Bastiaan Kleijn*[1,2]

[1]Sound and Image Processing Lab, School of Electrical Engineering,
KTH-Royal Institute of Technology, Stockholm, Sweden
[2]School of Engineering and Computer Science, Victoria University of Wellington,
Wellington, New Zealand

ABSTRACT

A non-intrusive quality predictor constitutes a mapping from signal features to a (typically one dimensional) representation of the perceived quality. Assuming that the regression model performing the mapping is suited to the data, the performance of the predictor largely depends on how well the parameters of this regression model can be inferred from the training data. In situations where the training data is scarce, model performance is degraded due to over-fitting. The effects of over-fitting can be mitigated by feature selection but the model performance remains low due to the insufficiently representative training data. The objective we pursue is to enhance the performance of a quality predictor by augmenting the feature set with the output of a pre-trained quality predictor. This approach introduces an implicit dependence of the regression model parameters on a larger amount of training data. In view of the increasing usage of speech signals with higher bandwidth, and the dearth of training data for such signals, an augmentation of particular interest is that of a wide-band feature set with a narrow-band quality prediction. Experimental results for additive noise and non-linear distortions encountered in hearing aids, using quality labels from an intrusive quality predictor, illustrate the performance enhancement capabilities of the proposed approach.

Index Terms— Non-intrusive quality assessment, machine learning with Gaussian processes, input uncertainty

1. INTRODUCTION

The objective evaluation of the perceptual quality of acoustic signals, based on signal features, is an integral component in the design and exploitation chains for speech and audio processing algorithms. Two classes of quality assessment models are commonly identified depending on the availability of a clean signal reference. *Intrusive* models, e.g., [1, 2] assume that a clean signal reference is available, while *non-intrusive* models, e.g., [3, 4] do not. The availability of the clean signal allows for the use of distance measures between the degraded and the reference signal. Mapping of such distance measures to quality predictions is, generally, more reliable compared to the mapping from signal features. Consequently, intrusive models achieve higher performance than non-intrusive models but are limited to off-line applications. In on-line applications, a clean signal reference is not available, which makes them the target domain for non-intrusive models. In this paper we consider the case of non-intrusive quality assessment for speech.

At a high level, the architecture of a non-intrusive quality assessment model consists of three processing stages as illustrated in Figure 1. The input signal $\mathbf{s}$ is first preprocessed to ensure the conformity between the test data and the data used for training the model. In practice, the signal is normalized to ensure a consistent level of presentation. Particular features $\mathbf{x}$ containing quality cues are then extracted from the normalized signal. Finally, a regression model maps the features into a quality indicator q.

Fig. 1. Non-intrusive QA system architecture.

Assuming that the signal is properly normalized, the two main factors that determine the performance of the quality predictor are i) the sensitivity of the feature set to signal variations for the considered distortion conditions and ii) the availability of training data so that given a particular regression model, the parameters of the latter can be inferred accurately. In practice, training data is costly and time consuming to collect as it involves the set-up of controlled subjective evaluation experiments. An additional obstacle is that such databases are mostly proprietary, which impedes their distribution.

In relation to the database shortage problem, we seek to enhance the performance of a non-intrusive quality predictor by augmenting its feature set with a prediction from a pre-trained model. We focus on the scenario where the performance of a wide-band (signal $\mathbf{s}$ sampled at 16 kHz) predictor is enhanced by augmenting its feature set with the prediction of a narrow-band ($\mathbf{s}$ sampled at 8 kHz) model trained on the same type of distortion conditions. The underlying assumption is that narrow-band (NB) and wide-band (WB) predictions are correlated. This approach can also be applied to the case where the sampling frequency remains the same. We shall refer to the predictor providing the augmentation feature as the auxiliary system and to the predictor providing the target quality estimate as the main system. The same characterization is used for the data available for training each of the two systems.

The feature set augmentation approach considered here is qualitatively different from the use of informative priors on the model parameters in a Bayesian inference context [5]. The auxiliary data in

The project LISTA acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 256230.

the latter case would be used to obtain a posterior distribution over the unknown model parameters. The posterior distribution would then be used as an informative prior when training with the main data. This set-up, however, requires that both systems have the same architectures, which doesn't hold in general and in particular for the case where the sampling frequencies differ. Furthermore, it is possible that the designer of the main system only observes the output from the pre-trained auxiliary system.

A secondary objective of our work is to verify that the variance of the prediction of the auxiliary system, assuming that such is available, can be used to improve the performance of the main system. From a modeling perspective, it is expected that the more general model, where the input uncertainty is taken into account, should perform at least as well as the simpler model. This expectation is conditional on the availability of a sufficient amount of data for training the more general model.

A suitable framework for implementing and validating these ideas is presented by Gaussian processes (GP) [6], and in particular GP with uncertain inputs [7]. Previous work on the topic [8] produced promising results. Here we extend this work by a more detailed analysis of the proposed method, addressing the remaining technical issues related to estimating the model parameters and performing a broader validation that also includes a comparison between the augmentation approaches with and without consideration for the input uncertainty. The data used for training and validation is synthetically generated using an intrusive quality assessment model operating in narrow-band and wide-band modes [1]. We considered a range of additive noise and non-linear distortion conditions representative of the distortions observed in non-linear hearing aids [9, 10].

The remainder of this paper is organized as follows. Section 2 describes the architectures of the considered systems. Section 3 introduces the theoretical framework on which we base the implementation of the systems. Section 4 presents the experimental validation set-up and the results. Conclusions are given in Section 5.

2. MODEL ARCHITECTURES

This section presents the model architectures used to address the objectives of this paper. The latter were identified as i) improving the performance of a wide-band non-intrusive quality predictor by augmenting its feature set with a narrow-band quality prediction from a pre-trained model and ii) increasing the performance gain further by taking into account the output variance of the pre-trained predictor. There are, therefore, three alternative system architectures whose performance we shall compare.

The augmentation of the feature set of a wide-band predictor with a narrow-band prediction, can be expected to produce a gain in performance when the NB and the WB quality ratings are correlated, and the amount of WB training data is limited. The requirement for correlation between the two quality estimates is typically satisfied for a wide range of distortion conditions as illustrated in Figure 2. The conditions included there consist of additive noise and non-linear signal distortions characteristic of non-linear hearing aids [9, 10]. The quality ratings were obtained with an intrusive quality assessment model [1].

The three system architectures are illustrated in Figure 3. Diagram A represents the baseline system, i.e., the system whose feature set is not augmented. Diagrams B and C depict similar systems that both adopt the feature set augmentation approach. The fundamental difference is that the system from diagram C accounts for the variance of the NB predictor in the regression model of the WB predictor

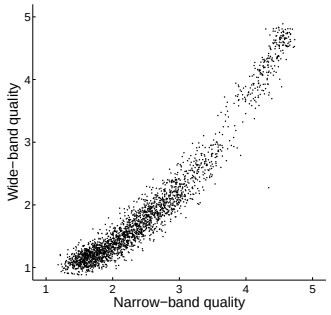

Fig. 2. Scatter plot over narrow-band and wide-band quality ratings (as obtained with an intrusive quality assessment model [1]).

while the system from diagram B does not. This is indicated in Figure 3, where the mean and the variance of the stochastic variable q_{NB} are denoted by $\mathrm{E}\left[\mathrm{q}_{\mathrm{NB}}\right]$ and $\mathrm{var}\left[\mathrm{q}_{\mathrm{NB}}\right]$ respectively.

The outputs of the three systems are denoted by $\mathrm{q}_{\mathrm{WB,A}}$, $\mathrm{q}_{\mathrm{WB,B}}$ and $\mathrm{q}_{\mathrm{WB,C}}$. We assume that the quality prediction is continuous in nature. While in practice perceptual tests for subjective quality assessment may use a discrete rating scale, e.g., [11], quality predictors are typically trained on mean opinion scores (MOS) [3, 4, 12] obtained by averaging the subjective responses over a large group of individuals. The motivation for doing this is to smooth out the individual preferences and obtain an estimate of the average perceived quality.

The databases used to train the NB and the WB predictors are denoted by $\mathbf{D}_{\mathrm{NB}}$ and $\mathbf{D}_{\mathrm{WB}}$ respectively. $\mathbf{x}_{\mathrm{NB}}$ and $\mathbf{x}_{\mathrm{WB}}$ in turn, represent the NB and the WB feature sets for signal $\mathbf{s}$.

The feature set $\mathbf{x}_{\mathbf{WB}}$ can be computed independently of $\mathbf{x}_{\mathrm{NB}}$. Alternatively, $\mathbf{x}_{\mathrm{NB}}$ can be treated as a subset of $\mathbf{x}_{\mathrm{WB}}$. The subset approach presents an elegant way to reduce the computational complexity of the quality predictor. One possible way to achieve this, is to extract the features from the bands of an auditory filter-bank [4, 13]. A point of practical significance is that a two-fold increase in the signal sampling frequency results in a less than two-fold increase in the number of auditory channels [13] and leads to a more compact feature set.

The signal analysis in non-intrusive quality assessment models is performed on a per-frame basis, e.g., [3, 4, 12, 14]. In [4] the per-frame features are mapped into an intermediate quality representation followed by aggregation to obtain the overall quality prediction. The more common approach [3, 12, 14] is to first obtain a global set of features and using a single mapping obtain the quality prediction. Here we adopt the latter approach using the features set from [15].

To obtain a fair comparison of the three model architectures presented in Figure 3, it is important that the main predictors, i.e., the predictors that output the target quality prediction share a common mathematical framework. Such a framework is obtained in terms of Gaussian processes [6] and a generalization thereof [7] to the case with input uncertainty. Relevant theoretical aspects of Gaussian processes are presented in the following section.

3. THEORETICAL FRAMEWORK OF THE REGRESSION MODELS

We assume a general statistical model of the form

$$\mathrm{q} = f\left(\mathbf{x}\right) + \epsilon, \;\; \epsilon \sim N\left(0,\, \sigma^2\right), \tag{1}$$

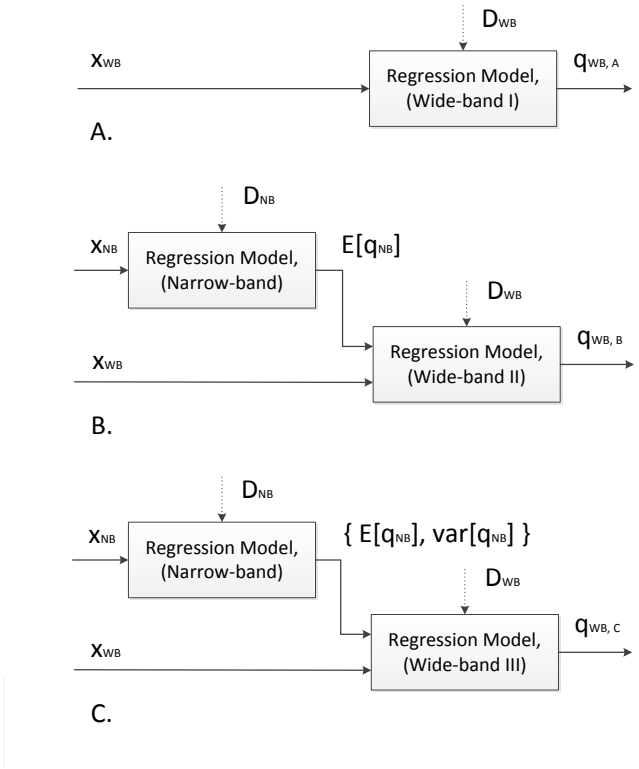

Fig. 3. System architectures: baseline (left), deterministic augmentation (middle) and stochastic augmentation (right).

where the continuous variable q represents the overall quality of the signal, $\mathbf{x} = [\mathrm{x}_1, \cdots, \mathrm{x}_L]$ are the corresponding features and ϵ is an independent noise term reflecting the influence of factors that are not accounted for by the model $f(\cdot)$. Under the Gaussian process framework, a Gaussian prior is placed over the function space

$$f(\mathbf{x}) \sim GP\left(m(\mathbf{x}),\, k\left(\mathbf{x},\, \mathbf{x}'\right)\right), \tag{2}$$

$$m(\mathbf{x}) = \mathrm{E}\left[f(\mathbf{x})\right] \tag{3}$$

$$k\left(\mathbf{x},\, \mathbf{x}'\right) = \mathrm{E}\left[\left(f(\mathbf{x}) - m(\mathbf{x})\right)\left(f\left(\mathbf{x}'\right) - m\left(\mathbf{x}'\right)\right)\right]. \tag{4}$$

If the process, which is indexed by $\mathbf{x}$, is stationary, its mean is constant. Assuming this to hold, we set $m(\mathbf{x}) = 0$ and adjust the mean of the observations by centering them around zero. The Gaussian process is then fully defined by its covariance function. A prerequisite for good modeling performance is that the covariance function reflects the properties of the data.

The joint distribution of the quality scores $\mathbf{q} = [\mathrm{q}_1, \cdots, \mathrm{q}_n]^{\mathrm{T}}$, from the training database, where n is the number of signals in this database, and the function value $f(\mathbf{x}_*)$ for a test signal $\mathbf{x}_*$, under the prior from equation (2), is

$$\begin{bmatrix} \mathbf{q} \\ y_* \end{bmatrix} \sim N\left(\mathbf{0},\; \begin{bmatrix} K(\mathbf{X},\, \mathbf{X}) + \sigma^2 I & K(\mathbf{X},\, \mathbf{x}_*) \\ K(\mathbf{x}_*,\, \mathbf{X}) & K(\mathbf{x}_*,\, \mathbf{x}_*) \end{bmatrix}\right), \tag{5}$$

where $\mathbf{X}$ is a matrix in which row i contains the features $\mathbf{x}_i$ for signal i, $K(\cdot,\cdot)$ represents the covariance matrix of its two arguments where individual covariance components are given by $k(\cdot,\cdot)$ and $y_* = f(\mathbf{x}_*)$. The predictive distribution for the test signal is obtained as the conditional distribution $\mathrm{p}\left(y_* | \mathbf{X},\, \mathbf{q},\, \mathbf{x}_*\right)$ for which a closed-form solution exists [6].

3.1. Covariance functions

We chose the squared exponential covariance function with automatic relevance determination (ARD) [6] to address the cases without input uncertainty, i.e., the auxiliary predictors from System B and C, and the main predictors from System A and B in Figure 3. This covariance function is defined as

$$k\left(\mathbf{x},\, \mathbf{x}'\right) = \nu e^{-\frac{1}{2}\sum_{l=1}^{L} \mathrm{w}_l^{-1}\left(\mathrm{x}_l - \mathrm{x}'_l\right)^2}, \tag{6}$$

where ν is a global scale parameter, L is the dimensionality of the feature space and w_l is the relevance weight for feature dimension l. The choice of this covariance function can be motivated as follows. First, it imposes a smooth decrease of the covariation between observations in the feature space with the increase in the distance between them. This is a characteristic that, we expect, reflects the properties of quality assessment data. Second, the squared exponential covariance function facilitates analytical approximations for the case with input uncertainty [7]. Third, ARD results in soft feature selection. This is generally important when working with high-dimensional feature spaces and particularly suitable to our application considering that the feature set augmentation approach increases the dimensionality of the feature space.

Let one of the dimensions of the feature set be contaminated by a Normally-distributed noise. Without loss of generality we take this to be the last dimension:

$$\begin{aligned} \mathrm{x}_L &= \mathrm{u} + \xi,\;\; \xi \sim N\left(0,\, \delta^2\right), \\ \mathrm{x}'_L &= \mathrm{u}' + \xi',\;\; \xi' \sim N\left(0,\, \delta'^2\right), \end{aligned} \tag{7}$$

where $\mathrm{u} = \mathrm{E}\left[\mathrm{x}_L\right]$. This scenario relates to the situation where the deterministic wide-band feature set $\mathbf{x}_{\mathbf{WB}}$ is augmented with a stochastic narrow-band prediction, and is addressed by the main predictor of System C in Figure 3. We assume ξ and ξ' to be independent. In the presence of input uncertainty the Gaussian process prior does not hold any more.

An analytical approximation framework for Gaussian processes in the presence of input uncertainty is established in [7]. The main idea is to approximate the resulting process with a Gaussian process that preserves the mean and the covariance of the original process. The case with a single uncertain input can be treated as a special case of this framework.

According to the law of iterated expectations

$$\mathrm{E}_{\mathbf{x}}\left[\mathrm{E}\left[y|\mathbf{x}\right]\right] = \mathrm{E}\left[y|\mathbf{u}\right] = 0, \tag{8}$$

where $\mathbf{u} = [\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_{L-1}, \mathrm{u}]$, and we used that the process is zero-mean, i.e., $\mathrm{E}\left[y|\mathbf{x}\right] = 0$. The law of conditional variances states that

$$\begin{aligned} \mathrm{E}_{\mathbf{x}}\left[\mathrm{var}\left[y|\mathbf{x}\right]\right] + \mathrm{var}_{\mathbf{x}}\left[\mathrm{E}\left[y|\mathbf{x}\right]\right] &= \mathrm{E}_{\mathbf{x}}\left[\mathrm{var}\left[y|\mathbf{x}\right]\right] \\ &= \mathrm{var}\left[y|\mathbf{u}\right]. \end{aligned} \tag{9}$$

The result from equation (9) extends to the covariances as:

$$\begin{aligned} \mathrm{E}_{\mathbf{x}}\left[\mathrm{cov}\left[y,\, y'|\mathbf{x},\, \mathbf{x}'\right]\right] &= \mathrm{E}_{\mathbf{x}}\left[k\left(\mathbf{x},\, \mathbf{x}'\right)\right] \\ &= \mathrm{cov}\left[y,\, y'|\mathbf{u},\, \mathbf{u}'\right] \\ &= \tilde{k}\left(\mathbf{u},\, \mathbf{u}'\right). \end{aligned} \tag{10}$$

We obtain the new covariance function by applying the expectation operator in equation (10) of the form:

$$\begin{aligned}\tilde{k}\left(\mathbf{u},\,\mathbf{u}'\right) &= \nu\sqrt{\frac{\mathrm{w}_L}{\mathrm{w}_L+\delta^2+\delta'^2}}e^{-\frac{1}{2}\sum_{l=1}^{L}\tilde{\mathrm{w}}_l^{-1}\left(\mathrm{u}_l-\mathrm{u}'_l\right)^2}\\ \tilde{\mathrm{w}}_l &= \mathrm{w}_l,\;\; l\in\{1,\cdots,L-1\}\\ \tilde{\mathrm{w}}_L &= \mathrm{w}_L+\delta^2+\delta'^2. \end{aligned} \tag{11}$$

The solution for the general case where all input dimensions are contaminated by Gaussian noise is provided in [7]. The variances δ^2 and δ'^2, first introduced in the equations from (7), represent the uncertainty in the output of the auxiliary (NB) quality predictor for the signals represented by the feature sets $\mathbf{x}_{\mathrm{NB}}$ and $\mathbf{x}'_{\mathrm{NB}}$ respectively.

We note that in the absence of input uncertainty, i.e., $\delta^2=\delta'^2=0$, the covariance function $\tilde{k}\left[\mathbf{u},\,\mathbf{u}'\right]$ from equation (11) converges to $k\left[\mathbf{x},\,\mathbf{x}'\right]$ from equation (6). In the presence of uncertainty, the contribution of the stochastic feature dimension is weighted such that its relevance decreases with the increase in the input variances. The overall covariance, in relation to the case without input uncertainty, is weighted down by a factor that reflects the relative change in the relevance of the stochastic dimension due to its inherent uncertainty.

3.2. Estimation of the model parameters

According to the prior placed on the observations, i.e., the mean quality ratings from the training database, the latter are jointly Normally distributed. We estimate the parameters $\left\{\mathrm{w}_1,\cdots,\mathrm{w}_L,\nu,\sigma^2\right\}$ of a model by minimizing the negative log-likelihood of these observations. The same approach is used for any of the auxiliary and the main systems, however, the covariance function is adapted to the particular case. Let

$$\mathbf{C}=K\left(\mathbf{X},\,\mathbf{X}\right)+\sigma^2 I. \tag{12}$$

The objective function is defined as:

$$\mathcal{O}=\frac{1}{2}\mathbf{q}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{q}+\frac{1}{2}\log\left|\mathbf{C}\right|+\frac{\mathrm{N}}{2}\log\left(2\pi\right). \tag{13}$$

where N is the number of signals in the training database.

To avoid the need for optimization constraints related to the positivity of the model parameters, and to facilitate the derivation, we adopt the common parametrization, e.g., [6],

$$\mathbf{r}=\left[\log\sqrt{\mathrm{w}_1},\cdots\log\sqrt{\mathrm{w}_L},\log\sqrt{\nu},\log\left(\sigma\right)\right]. \tag{14}$$

Representing the model parameters by $\mathbf{r}=[\mathrm{r}_1,\cdots,\mathrm{r}_{L+2}]$, the derivative for some r_m is given by

$$\begin{aligned}\frac{\partial\mathcal{O}}{\partial\mathrm{r}_m} &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\Omega\frac{\partial\mathbf{C}}{\partial\mathrm{r}_m}\right]_{i,j}\\ \Omega &= \mathbf{C}^{-1}-\mathrm{a}\mathrm{a}^{\mathrm{T}}\\ \mathrm{a} &= \mathbf{q}^{\mathrm{T}}\mathbf{C}^{-1}, \end{aligned} \tag{15}$$

where i and j index the respective rows in $\mathbf{X}$.

The covariance function $k\left(\mathbf{x},\,\mathbf{x}'\right)$ constitutes a well-studied case and is handled by most computational packages for Gaussian processes, e.g., [6]. The derivatives of the covariance function $\tilde{k}\left(\mathbf{u},\,\mathbf{u}'\right)$ with respect to the relevant model parameters were obtained as follows:

$$\frac{\partial\tilde{k}\left(\mathbf{u},\mathbf{u}'\right)}{\partial\log\sqrt{\nu}}=2\tilde{k}\left(\mathbf{u},\mathbf{u}'\right), \tag{16}$$

$$\begin{aligned}\frac{\partial\tilde{k}\left(\mathbf{u},\mathbf{u}'\right)}{\partial\log\sqrt{\mathrm{w}_l}} &= \tilde{k}\left(\mathbf{u},\mathbf{u}'\right)\mathrm{w}_l^{-1}\left(\mathrm{u}_l-\mathrm{u}'_l\right)^2,\\ &\quad l\in\{1,\cdots,L-1\}, \end{aligned} \tag{17}$$

$$\begin{aligned}\frac{\partial\tilde{k}\left(\mathbf{u},\mathbf{u}'\right)}{\partial\log\sqrt{\mathrm{w}_L}} &= \frac{\tilde{k}\left(\mathbf{u},\mathbf{u}'\right)}{\mathrm{w}_L+\delta^2+\delta'^2}\\ &\quad\left[\delta^2+\delta'^2+\frac{\mathrm{w}_L\left(\mathrm{u}_L-\mathrm{u}'_L\right)^2}{\mathrm{w}_L+\delta^2+\delta'^2}\right]. \end{aligned} \tag{18}$$

We note that the objective function $\mathcal{O}$ is non-convex and local optimality of the solution can only be assumed. This poses a problem when comparing the performance of the main predictors in the baseline and the augmented systems from Figure 3. We address the problem by careful selection of the initial point of the optimization process. This topic is discussed further in Section 4.1.

4. EXPERIMENTAL VALIDATION

The experimental validation procedure and the results thereof are presented in this section. Relevant details of the implementation are provided in Section 4.1. The data preparation is discussed in Section 4.2 followed by a description of the experimental set-up and presentation of the results in Section 4.3.

4.1. Implementation

The popularity of Gaussian processes for addressing regression and classification problems, e.g., [6, 16], has lead to a number of computational packages with high level of modularity allowing for the plug-in of covariance functions not included in the original distributions. We used the package accompanying [6], adding a covariance and derivative computation routine based on equations (11), (16), (17) and (18). This package uses a Polack-Ribiere form of the conjugate gradient approach [17] for computing the search direction during optimization.

In Section 3.2 we noted that the initialization of the optimization procedure is important due to the non-convexity of the objective function. We adopt the following strategy. To avoid confusion, let L denote the number of features in the augmented feature sets (the main predictors in Systems B and C). The feature set of the main predictor in system A, then contains $L-1$ features. The weight coefficients for the latter, denoted by $\mathbf{w}_{\mathrm{A}}$, are initialized using a flat-start approach ($\mathrm{w}_{\mathrm{A,l}}=\mathrm{w}_0,\; l\in\{1,\cdots,L-1\}$), i.e., all features are given equal a-priori weights. The value of w_0 is not critical due to the presence of the global scale parameter ν. The flat-start approach is also applied to the weights in the auxiliary predictors from systems B and C.

Initialization for the weight coefficients from the main predictor in system B, denoted by $\mathbf{w}_{\mathrm{B}}$, does not use a flat start. The general idea is to look for a solution in the vicinity of the solution for system A. The initial value for the coefficients in $\mathbf{w}_{\mathrm{B}}$ corresponding to the WB feature set before augmentation is the solution for $\mathbf{w}_{\mathrm{A}}$. The initial value for the weight of the augmentation feature, in this case, was experimentally shown not to be critical. For consistency, we use $\mathrm{w}_{\mathrm{B},L}=\mathrm{w}_0$.

Initialization for the weight coefficients from the main predictor in system C, denoted by $\mathbf{w}_{\mathrm{C}}$, is based on the solution for $\mathbf{w}_{\mathrm{B}}$. This allows us to refine the solution for the main predictor in System B

further by taking into account the stochastic nature of the augmentation feature.

The solution to the optimization process for the parameters of the main predictor in System C is sensitive to the initialization for the weight of the augmentation feature. Equation (18) suggests that increasing the value for $\log(\mathrm{w}_L)$ drives the respective derivative to zero. The use of a gradient optimizer, such as the one adopted in our implementation, can lead to a solution that renders the augmentation feature unimportant and converges, in theory, to the solution for the system without augmentation of the feature set. Random initialization for $\mathbf{w}_\mathrm{C}$ is, therefore, not a suitable approach. The progressive initialization strategy outlined above avoids this problem.

4.2. Data Preparation

We perform the validation of the feature set augmentation approach using a synthetically-labeled database with additive noise and non-linear distortions characteristic of non-linear hearing aids [9, 10]. These include one clean and 28 noisy conditions grouped into seven sets with four levels each:

1. Additive stationary speech-shaped noise at 5, 10, 15 and $20\,\mathrm{dB}$ signal-to-noise ratio (SNR);
2. Additive babble noise at 5, 10, 15 and $20\,\mathrm{dB}$ signal-to-noise ratio (SNR);
3. Scalar quantization using 4, 5, 6 and 7 bits;
4. Compression with a combination of different compression ratios, attack times and release times;
5. Babble noise and compression;
6. Babble noise and spectral subtraction;
7. Babble noise, spectral subtraction and compression.

The source material was taken from the P.Sup23 database [11] amounting to 832 signals. In generating the database, each condition was represented by 80 utterances resulting in a total of 2320 signals. As a result, each utterance was used, at most, three times and always with different distortion conditions. Signal normalization, as illustrated in Figure 1, was performed on the wide-band signals, thus, ensuring proper normalization for the narrow-band signals as well. The intrusive model from [1] was used to create the synthetic quality labels $\mathbf{q}_\mathrm{NB}$ and $\mathbf{q}_\mathrm{WB}$ as it is a popular benchmark and doesn't have reported issues with the considered distortions conditions.

The derivation of the global feature set is described in detail in [8]. In short, the per-frame features are computed from the log-domain band-powers in the channels of a modulation spectrum filter-bank, which operates on the output of an auditory filter-bank. The per-frame features are then converted to per-signal (global) features by computing the first and second order central moments of the individual per-frame features in speech active signal frames over the duration of the signal. Using an auditory filter-bank with 23 channels in the narrow-band and 30 channels in the wide-band case results in $\mathbf{x}_\mathrm{NB} \in \mathbb{R}^{92}$ and $\mathbf{x}_\mathrm{WB} \in \mathbb{R}^{120}$ respectively. To avoid numerical issues during optimization as well as bias to features based on their numerical scale, the global features are normalized over the training database to have a mean of zero and a standard deviation of one.

4.3. Experimental Results

Experiments were performed using a five-round cross-validation procedure to improve the significance of the results. In each round the data were split into three sets: $1/5$ of the data were used for validation, $2/5$ were used as $\mathbf{D}_\mathrm{NB}$ to train the auxiliary (NB) predictor and $2/5$ were used as $\mathbf{D}_\mathrm{WB}$. Five operating points were considered with respect to the fraction of $\mathbf{D}_\mathrm{WB}$ that was used for training the main (WB) predictors. These were $1/5$, $2/5$, $3/5$, $4/5$ and all (1) of $\mathbf{D}_\mathrm{WB}$. In terms of all the available data these fractions are: $2/25$, $4/25$, $6/25$, $8/25$ and $2/5$.

In each round of the cross-validation and for each wide-band training data configuration we computed the values of four established measures for assessing the performance of quality predictors: Pearson correlation coefficients per-file (ρ_pf) and per-condition (ρ_pc), and root-mean-squared (RMS) error per-file (r_pf) and per-condition (r_pc) [1, 3]. A third-order monotonic polynomial mapping between predictions and targets was performed for consistency with established evaluation procedures [1, 3].

The results averaged over the five rounds of the cross-validation are presented graphically in Figure 4. We note that for all four measures, the baseline system is outperformed by both systems using the augmentation approach. As expected, the performance difference decreases as the amount of wide-band training data increases. This can be explained with the relative decrease in the contribution of the narrow-band prediction to modeling the wide-band prediction.

Figure 4 can also be used to compare the performance of the two systems using augmented feature sets. We note that, on average, and with the exception of the $1/5$ wide-band training data configuration, the system with input uncertainty outperforms its counterpart. The results for the $1/5$ configuration do not indicate that it is better to ignore the stochastic nature of the augmentation feature. The regression model based on the covariance function from (11) generalizes the model based on (6). At the $1/5$ configuration, the absolute amount of data is insufficient to train the more general model well. Note that for this configuration less than 200 data points are available to train a model with 123 parameters.

To gain an insight into the performance differences between the two systems using feature set augmentation, we performed a statistical significance test. For each performance measure and wide-band training data configuration we tested two hypotheses. The first hypothesis was that the performance figures for System A and System B come from the same distribution. The second hypothesis was that the performance figures for System A and System C come from the same distribution. The two series of values on which the significance test operates, thus, contain five members each, one for every round of the cross-validation. Avoiding any assumptions on the distribution of the population in the series, we used the Wilcoxon signed rank test [18]. In line with the results from Figure 4, the improvement was more significant for system C in all cases except the one with the minimum amount of WB training data.

5. CONCLUSIONS

Feature set augmentation for one quality predictor with the output of another, where the second system is pre-trained, possibly using a large amount of training data, can be used to improve the performance of the quality assessment system, especially when the amount of data for training it is limited. This approach can be applied when the quality predictors operate at the same sampling frequency, but also when the sampling frequencies differ. This is an attractive possibility in view of the gradually increasing sampling rates of audio processing algorithms and the limited amount of training data. We have shown, using data with quality labels generated with an intrusive quality assessment model, that feature set augmentation for a wide-band quality predictor with the output from a narrow-band quality predictor improves the performance of the wide-band predictor. As expected, the improvement is most significant when the data

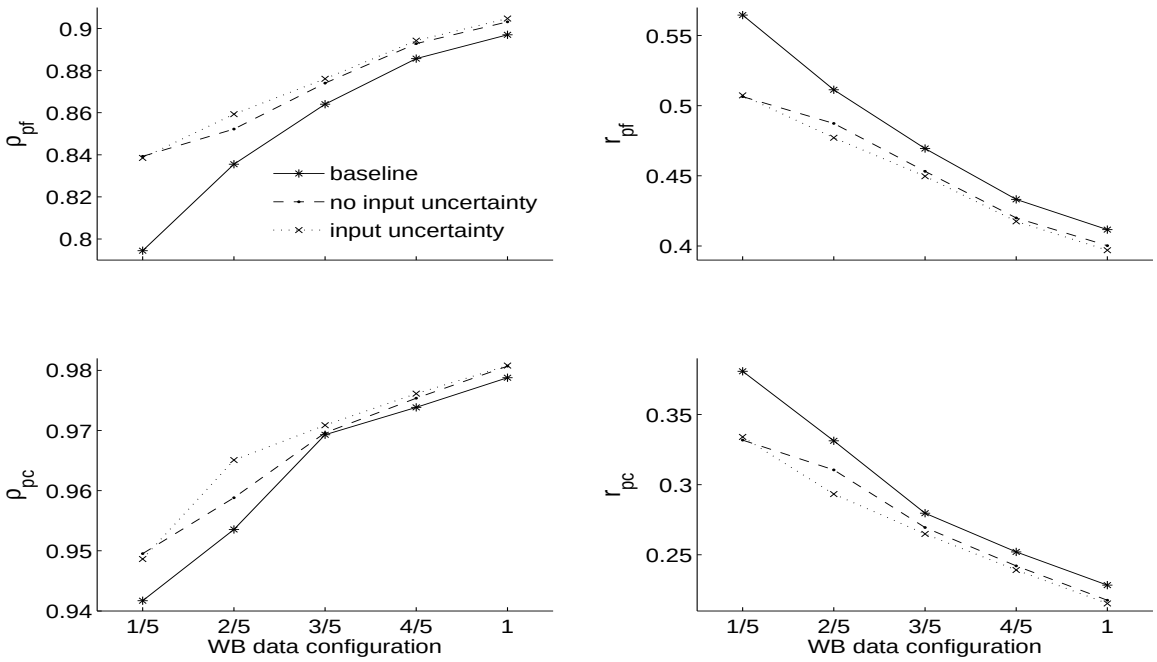

Fig. 4. Performance results: Pearson correlation coefficient per-file (ρ_{pf}) and per-condition (ρ_{pc}), and root-mean-squared error per-file (r_{pf}) and per-condition (r_{pc}) for the baseline and the two augmented-feature-set systems.

available for training the wide-band predictor is limited. Taking into consideration the variance of the narrow-band prediction (if such is available) enhances further the performance gain from the augmentation approach. The experimental results outline a bound-like behavior for the performance of the augmented-feature-set predictors relative to the case without augmentation.

References

[1] ITU-T Rec. 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," 2001.

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

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

[4] D. S. Kim and A. Tarraf, "ANIQUE+: A New American National Standard for Non-intrusive Estimation of Narrowband Speech Quality," *Bell Labs Tech. Journal*, vol. 12, pp. 221–236, 2007.

[5] A. Gelman, J. B. Carlin, H.S. Stern, and D. B. Rubin, *Bayesian Data Analysis, 2nd Edition*, Chapman & Hall, 2009.

[6] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[7] A. Girard, *Approximate Methods for Propagation of Uncertainty with Gaussian Process Models*, Ph.D. thesis, University of Glasgow, 2004.

[8] P. N. Petkov and W. B. Kleijn, "Objective Quality Estimation of Wide-Band Speech Using a Narrow-Band Prior," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, 2010.

[9] K. H. Arehart, J. M. Kates, M. C. Anderson, and L.O. Harvey, "Effects of Noise and Distortion on Speech Quality Judgments in Normal-Hearing and Hearing-Impaired Listeners," *J. Acoust. Soc. Am.*, vol. 122, pp. 1150–1164, 2007.

[10] J. M. Kates and K. H. Arehart, "A Time-Frequency Modulation Model of Speech Quality," in *Proc. IEEE Workshop Appl. Sig. Process. Audio Acoust.*, 2007, pp. 231–234.

[11] ITU-T Rec. P.Sup23, "ITU-T Coded-Speech database," 1998.

[12] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-Complexity, Nonintrusive Speech Quality Assessment," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 14, pp. 1948–1956, 2006.

[13] B. C. Moore, *An Introduction to the Psychology of Hearing*, Elsevier Academic Press, 2004.

[14] T. H. Falk and W. Y. Chan, "Nonintrusive Speech Quality Estimation Using Gaussian Mixture Models," *IEEE Sig. Proc. Letters*, vol. 13, pp. 108–111, 2006.

[15] P. N. Petkov, I. S. Mossavat, and W. B. Kleijn, "A Bayesian Approach to Non-Intrusive Quality Assessment of Speech," in *Proc. Interspeech*, 2009, pp. 2875–2878.

[16] D. J. C. MacKay, "Introduction to Gaussian Processes," in *Neural Networks and Machine Learning*, C. M. Bishop, Ed., vol. 168 of *NATO ASI*. Springer, Berlin, 1998.

[17] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, 1999.

[18] D. F. Bauer, "Constructing Confidence Sets Using Rank Statistics," *J. Am. Stat. Assoc.*, vol. 67, pp. 687–690, 1972.