Nonlinear Perception of Hearing-Impaired People Using Preference Learning with Gaussian Processes

P.C. Groot[1], Tom Heskes[1], and Tjeerd M.H. Dijkstra[1,2,3]

[1] Radboud Univ. Nijmegen, Intelligent Systems, the Netherlands

[2] GN ReSound, Algorithm R&D dept. Eindhoven, the Netherlands

[3] Technical University Eindhoven, Department of Electrical Engineering, the Netherlands

perry@cs.ru.nl, tomh@cs.ru.nl, t.dijkstra@cs.ru.nl

Abstract

In this report, we describe a probabilistic kernel approach to pairwise preference learning based on Gaussian Processes proposed in Chu and Ghahramani [2005] and apply the method to audio input signals. The objective is to have a principled approach for modeling and predicting speech quality for arbitrary degradation mechanisms that might be present in a hearing aid. In Arehart et al. [2007] pairwise comparisons were performed with 14 normal-hearing and 18 hearing-impaired subjects for several sound distortions. The kernel approach gives a significant improvement in the prediction of sound quality perception over previous results. We show a significant difference between normal-hearing and hearing-impaired subjects, because of nonlinearities in the perception of hearing-impaired subjects. In this report, we describe a probabilistic kernel approach to pairwise preference learning based on Gaussian Processes proposed in Chu and Ghahramani [2005] and apply the method to audio input signals. The objective is to have a principled approach for modeling and predicting speech quality for arbitrary degradation mechanisms that might be present in a hearing aid. In Arehart et al. [2007] pairwise comparisons were performed with 14 normal-hearing and 18 hearing-impaired subjects for several sound distortions. The kernel approach gives a significant improvement in the prediction of sound quality perception over previous results. We show a significant difference between normal-hearing and hearing-impaired subjects, because of nonlinearities in the perception of hearing-impaired subjects.

1 introduction

A central issue in the development of hearing aids or other communicating devices is the sound quality that is perceived by their users. The perceived quality is affected by noise present in the input signal as well as linear and nonlinear distortions that result from signal processing within the device itself [Souza et al., 2006, Stelmachowicz et al., 1999]. A number of methods have been developed in the last decades for measuring the perception of sound quality, including a multitone test signal with logarithmically spaced components [Czerwinski et al., 2001a,b], vowel sounds [Levitt et al., 1987], comb-filtered noise [Kates, 1990], and coherence based methods [Arehart et al., 2007, Dyrlund, 1992, Preves, 1990]. Tan and Moore have written several papers on the topic focusing on linear distortion [Moore and Tan, 2003, 2004], nonlinear distortion [Tan et al., 2003, 2004], and their combination [Tan and Moore, 2008]. Although some of these models have been developed using normal-hearing subjects only, prediction of sound quality perception was also found to be reasonable for hearing-impaired subjects [Arehart et al., 2007, Tan and Moore, 2008], but some systematic errors remain [Arehart et al., 2007] and some model extensions have been surprisingly ineffective possibly because of random variability in the judgments of subjects [Tan and Moore, 2008].

It is well-known, however, that hearing-impaired subjects appear to have only moderate test-retest reliability when judging sound quality [Gabrielsen et al., 1998, Narendran and Humes, 2003]

and consistency of experimental results suggest that hearing-impaired subjects are either less sensitive than normal to changes in nonlinear distortion or that there are greater individual differences between hearing-impaired subjects [Tan and Moore, 2008]. Nevertheless, current models for predicting perceived sound quality do not or can not make a clear distinction between both groups of subjects.

In Arehart et al. [2007] pairwise comparisons were performed with 14 normal-hearing and 18 hearing-impaired subjects for several sound distortions. Arehart et al. [2007] analyzed their data by (1) pooling responses over all normal-hearing listeners and second stimulus presentations, which resulted in a preference probability ($0 \leq p \leq 1$) for each of the 24 distortions (3 types and 8 levels each); (2) by regressing the preference probability on a three-level coherence based speech intelligibility (SII) measure, they obtained three regression coefficients (plus constant); (3) with a log-sigmoid function they transformed the fitted regression model into a quality metric termed Q_3. We extend their analysis by (1) fitting a model to individual listeners; (2) directly fitting the binary response data; (3) using a flexible *non-parametric* regression model using Gaussian Processes [Chu and Ghahramani, 2005, Rasmussen and Williams, 2006]; (4) devising model-independent measures for response bias and consistency. We show that the predictive performance for hearing-impaired subjects can be significantly improved by using their own individual preferences, taking into account a response bias and inconsistencies in user preferences, and allowing for nonlinearities in perception of hearing-impaired subjects. As no such improvements could be made for normal-hearing subjects, this demonstrates significant differences between the groups of normal-hearing and hearing-impaired subjects.

The rest of this paper is organized as follows. Section 2 describes the Bayesian framework using preference learning with GPs. Section 3 describes how to perform classification with GPs using a Laplace approximation. Section 4 describes a maximum likelihood approach to the model selection problem. Section 5 compares the classification results of Arehart et al. [2007] with a linear and nonlinear classifier based on the GP approach. Section 6 gives our conclusions. Notation and mathematical derivations are given in Appendices A–D.

2 Bayesian Framework

Let $X = \{x_i, \ldots, x_n\}$ be a set of n distinct instances (e.g., sound samples) with $x_i \in \mathbb{R}^d$ (e.g., sound features). Let $\mathcal{D}$ be a set of m observed pairwise preference comparisons over instances in X, i.e.,

$$\mathcal{D} = \{(v_{i,1}, v_{i,2}, d_i) \,|\, 1 \leq i \leq m, v_{i,1} \in X, v_{i,2} \in X, d_i \in \{-1, 1\}\} \tag{1}$$

where $d_i = 1$ when $v_{i,1} \succ v_{i,2}$ and $d_i = -1$ otherwise with $v_{i,1} \succ v_{i,2}$ meaning that $v_{i,1}$ is preferred over $v_{i,2}$. For example, the $v_{i,1}, v_{i,2}$ could be two sound samples (possibly related to some hearing aid parameter settings) and the user has to decide which sound sample he prefers.

The idea, is that there is an unobserved latent function $f(x_i)$ associated with each training sample x_i such that the function values $\{f(x_i)\}$ preserve the preference relations of the subject. We use a Gaussian Process prior on these latent function values. Together with an appropriate likelihood function, the latent functions can be learned in a Bayesian framework from pairwise preferences between samples.

2.1 Gaussian Process Prior

We assume that the latent function values $\{f(x_i)\}$ are a realization of random variables in a zero-mean Gaussian Process, which can be fully specified by a covariance matrix. An often used kernel is the Gaussian or Squared Exponential kernel:

$$K(f(x_i), f(x_j)) = \exp\left(-\frac{\kappa}{2}\sum_{l=1}^{n}(x_i^l - x_j^l)^2\right) \tag{2}$$

where $\kappa > 0$ and x_p^l is the l-th element or dimension of x_p. This kernel leads to the following multivariate Gaussian prior of latent function values $\{f(x_i)\}$

$$p(\boldsymbol{f}) = \frac{1}{(2\pi)^{\frac{n}{2}}|K|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{f}^T K^{-1}\boldsymbol{f}\right) \tag{3}$$

where $\boldsymbol{f} = [f(x_1), f(x_2), \ldots, f(x_n)]^T$ is a vector of function values on the n distinct instances in X. The matrix K is the $n \times n$ covariance matrix where each ij-th element is $K(f(x_i), f(x_j))$.

2.2 Likelihood $p(\mathcal{D}|\boldsymbol{f})$

The likelihood is the joint probability of observing the preferences given the latent function. We assume that the likelihood can be evaluated on individual observations:

$$p(\mathcal{D}|\boldsymbol{f}) = \prod_{k=1}^{m} p(v_{k,1}, v_{k,2}, d_k | f(v_{k,1}), f(v_{k,2})) \tag{4}$$

Here, we use a slightly modified version of the likelihood function proposed by Chu and Ghahramani [2005], which is defined as follows for ideally noise-free cases:

$$p_{ideal}(v_{k,1}, v_{k,2}, d_k | f(v_{k,1}), f(v_{k,2})) = \begin{cases} 1 & \text{if } f(v_{k,1}) > f(v_{k,2}) \text{ and } d_k = 1 \\ 1 & \text{if } f(v_{k,1}) \leq f(v_{k,2}) \text{ and } d_k = -1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

which requires that the latent function values $\{f(x_i)\}$ preserve the preference relations of the user.

Human users are, however, not always consistent in stating their preferences. To allow for noise in the preference relations we assume that the individual observations are contaminated with Gaussian noise, i.e., both $\delta_1, \delta_2 \sim \mathcal{N}(0, \sigma^2)$, hence $\delta_1 - \delta_2 \sim \mathcal{N}(0 - 0, \sigma^2 + \sigma^2) = \mathcal{N}(0, (\sqrt{2}\sigma)^2)$. Furthermore, we extend the likelihood function of Chu and Ghahramani [2005] by including a user response bias b that depends on the order of samples presented, with $b < 0$ denoting a response bias for the first sample, $b > 0$ denoting a response bias for the second sample, and $b = 0$ denoting no response bias for both samples. Both σ, b are so called nuisance parameters of the model. Then

$$\begin{aligned}
p(v_{k,1}, v_{k,2}, 1 | f(v_{k,1}), f(v_{k,2})) &= p(f(v_{k,1}) + \delta_1 > f(v_{k,2}) + b + \delta_2) \\
&= p(\delta_1 - \delta_2 > f(v_{k,2}) + b - f(v_{k,1})) \\
&= p(\delta_2 - \delta_1 < f(v_{k,1}) - f(v_{k,2}) - b) \\
&= \int_{-\infty}^{\boldsymbol{f}(v_{k,1}) - \boldsymbol{f}(v_{k,2}) - b} \frac{1}{(2\pi(\sqrt{2}\sigma)^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2(\sqrt{2}\sigma)^2}x^2\right) dx \\
&= \int_{-\infty}^{f(v_{k,1}) - f(v_{k,2}) - b} \frac{1}{(\sqrt{2}\sigma)} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2 \frac{1}{(\sqrt{2}\sigma)^2}\right) dx \\
&= \int_{-\infty}^{z_k} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) du \\
&= \Phi(z_k)
\end{aligned} \tag{6}$$

where we used the substitution $u = \frac{x}{\sqrt{2}\sigma}$, $du = \frac{1}{\sqrt{2}\sigma}$, and $z_k = \frac{f(v_{k,1}) - f(v_{k,2}) - b}{\sqrt{2}\sigma}$. Analogously, $p(v_{k,1}, v_{k,2}, -1) = \Phi(z_k)$ with $z_k = \frac{f(v_{k,2}) - f(v_{k,1}) + b}{\sqrt{2}\sigma}$. Summarizing

$$p(v_{k,1}, v_{k,2}, d_k | f(v_{k,1}), f(v_{k,2})) = \Phi(z_k) \text{ with } z_k = \frac{d_k(f(v_{k,1}) - f(v_{k,2}) - b)}{\sqrt{2}\sigma} \tag{7}$$

Other likelihood models, such as the Bradley-Terry-Luce (BTL) model [Bradley and Terry, 1952, Luce, 1959], can also easily be incorporated.

2.3 Posterior

Using Bayes' rule we can compute the posterior probability as follows

$$p(\boldsymbol{f}|\mathcal{D}) = \frac{p(\boldsymbol{f})p(\mathcal{D}|\boldsymbol{f})}{p(\mathcal{D})} = \frac{p(\boldsymbol{f})}{p(\mathcal{D})}\prod_{k=1}^{m}\Phi(z_k) \tag{8}$$

with the prior probability as defined in Eq. (3), the likelihood function as defined in Eq. (4), and the normalization constant $p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{f})p(\boldsymbol{f})d\boldsymbol{f}$.

The Bayesian framework as described here, depends on the model parameters including the kernel parameter κ controlling the kernel shape, and the likelihood parameters σ, b with σ controlling the noise level and b controlling the response bias. The parameters κ, σ, b are collected into θ, which we call the hyperparameter vector.

3 Gaussian Process Classification

In order to compute a predictive probability using the full Bayesian framework as described in Section 2, we need to integrate over the hyperparameters in θ-space. This integral is intractable, but several approaches can be followed to approximate the integral effectively. These approaches include sampling methods, such as Monte Carlo methods, or deterministic approximation methods, such as Laplace [MacKay, 1994] or Expectation Propagation [Minka, 2001]. Here, we consider the Laplace method, which approximates the posterior distribution $p(\boldsymbol{f}|\mathcal{D})$ as a Gaussian.

3.1 Laplace Approximation

The maximum a posteriori (MAP) estimate of the latent function values, denoted $\hat{\mathbf{f}}$, is the mode of the posterior distribution given in Eq. (8). As the MAP estimate is independent of the evidence $p(\mathcal{D})$ and the logarithm is a monotonic function, the following equality holds

$$\hat{\mathbf{f}} = \arg\max_{\boldsymbol{f}} p(\boldsymbol{f}|\mathcal{D}) = \arg\max_{\boldsymbol{f}}\{\ln p(\boldsymbol{f}) + \ln p(\mathcal{D}|\boldsymbol{f})\} \tag{9}$$

Using the Gaussian Process prior $p(\boldsymbol{f})$ defined in Eq. (3), define

$$\begin{aligned}\Psi(\boldsymbol{f}) &= \ln p(\mathcal{D}|\boldsymbol{f}) + \ln p(\boldsymbol{f}) \\ &= \ln p(\mathcal{D}|\boldsymbol{f}) + \ln \frac{1}{(2\pi)^{\frac{n}{2}}|K|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\boldsymbol{f}^T K^{-1}\boldsymbol{f}\right) \\ &= \ln p(\mathcal{D}|\boldsymbol{f}) - \frac{1}{2}\boldsymbol{f}^T K^{-1}\boldsymbol{f} - \frac{1}{2}\ln|K| - \frac{n}{2}\ln 2\pi\end{aligned} \tag{10}$$

Then differentiating w.r.t. $\boldsymbol{f}$ we obtain

$$\begin{aligned}\nabla\Psi(\boldsymbol{f}) &= \nabla \ln p(\mathcal{D}|\boldsymbol{f}) - K^{-1}\boldsymbol{f} \\ \nabla\nabla\Psi(\boldsymbol{f}) &= \nabla\nabla \ln p(\mathcal{D}|\boldsymbol{f}) - K^{-1} = -W - K^{-1} = -(K^{-1} + W)\end{aligned} \tag{11}$$

where we defined $W = -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{f})$. Taking the Laplace approximation of $\Psi(\boldsymbol{f})$ amounts to taking the second order Taylor expansion in the maximum a posteriori (MAP) estimate $\hat{\mathbf{f}}$, which gives (note that $\nabla\Psi(\hat{\mathbf{f}}) = 0$):

$$\begin{aligned}\Psi(\boldsymbol{f}) &\simeq \Psi(\hat{\mathbf{f}}) + \frac{1}{2}(\boldsymbol{f} - \hat{\mathbf{f}})^T\nabla\nabla\Psi(\hat{\mathbf{f}})(\boldsymbol{f} - \hat{\mathbf{f}}) \\ &= \Psi(\hat{\mathbf{f}}) - \frac{1}{2}(\boldsymbol{f} - \hat{\mathbf{f}})^T[-\nabla\nabla\Psi(\hat{\mathbf{f}})](\boldsymbol{f} - \hat{\mathbf{f}}) \\ &= \Psi(\hat{\mathbf{f}}) - \frac{1}{2}(\boldsymbol{f} - \hat{\mathbf{f}})^T[K^{-1} + W](\boldsymbol{f} - \hat{\mathbf{f}})\end{aligned} \tag{12}$$

This approximation can then be used to approximate the posterior distribution as a Gaussian with mean $\hat{\mathbf{f}}$) and covariance matrix given by the negative inverse Hessian of Ψ:

$$\begin{aligned} p(\boldsymbol{f}|\mathcal{D}) &\propto p(\boldsymbol{f})p(\mathcal{D}|\boldsymbol{f}) = \exp\Psi(\boldsymbol{f}) \\ &\simeq \exp\left(\Psi(\hat{\mathbf{f}}) - \frac{1}{2}(\boldsymbol{f}-\hat{\mathbf{f}})^T[K^{-1}+W](\boldsymbol{f}-\hat{\mathbf{f}})\right) \propto \mathcal{N}(\hat{\mathbf{f}}, (K^{-1}+W)^{-1}) \end{aligned} \tag{13}$$

3.2 Prediction

Given test samples X_* we have a prior joint multivariate Gaussian distribution, i.e.,

$$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{f}_* \end{bmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(X,X) & K(X,X_*) \\ K(X,X_*)^T & K(X_*,X_*) \end{pmatrix}\right] = \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K & K_* \\ K_*^T & K_{**} \end{pmatrix}\right] \tag{14}$$

where we introduced the abbreviations $K = K(X,X)$, $K_* = K(X,X_*)$, and $K_{**} = K(X_*,X_*)$. The conditional distribution $p(\boldsymbol{f}_*|\boldsymbol{f})$ is a Gaussian too (cf. [Rasmussen and Williams, 2006, Eq. (A6)]), i.e.,

$$p(\boldsymbol{f}_*|\boldsymbol{f}) \sim \mathcal{N}(K_*^T K^{-1}\boldsymbol{f}, K_{**} - K_*^T K^{-1} K_*) \tag{15}$$

The predictive distribution $p(\boldsymbol{f}_*|\mathcal{D})$ can be computed by integrating over $\boldsymbol{f}$-space

$$p(\boldsymbol{f}_*|\mathcal{D}) = \int p(\boldsymbol{f}_*|\boldsymbol{f})p(\boldsymbol{f}|\mathcal{D}) \tag{16}$$

where the posterior distribution $p(\boldsymbol{f}|\mathcal{D})$ can be approximated with a Gaussian using the Laplace approximation, i.e., (cf. [Rasmussen and Williams, 2006, Eq. (3.20)])

$$p(\boldsymbol{f}|\mathcal{D}) \sim \mathcal{N}(\hat{\mathbf{f}}, (K^{-1}+W)^{-1}) \tag{17}$$

where $W = -\nabla\nabla \ln p(D|\boldsymbol{f})$. Hence, the predictive distribution can be approximated by a convolution of two Gaussians which results in a Gaussian $\mathcal{N}(f_*; \mu^*, K^*)$ where

$$\begin{aligned} \mu^* &= K_*^T K^{-1}\hat{\mathbf{f}} \\ K^* &= K_{**} - K_*^T(W^{-1}+K)^{-1}K_* \end{aligned} \tag{18}$$

The predictive preference $p(v_{k,1}, v_{k,2}, 1|\mathcal{D})$ can be evaluated by the integral $\int p(v_{k,1}, v_{k,2}, 1|\boldsymbol{f}_*\mathcal{D})p(\boldsymbol{f}_*|\mathcal{D})d\boldsymbol{f}_*$, which is a Gaussian convoluted with the cumulative Gaussian and yields

$$p(v_{k,1}, v_{k,2}, 1|\mathcal{D}) = \Phi\left(\frac{\mu^*_{v_{k,1}} - \mu^*_{v_{k,2}} - b}{\sigma_*}\right) \tag{19}$$

where $\sigma_*^2 = 2\sigma^2 + K^*(v_{k,1}, v_{k,1}) + K^*(v_{k,2}, v_{k,2}) - K^*(v_{k,1}, v_{k,2}) - K^*(v_{k,2}, v_{k,1})$ with σ, b the hyperparameters of the likelihood function.

4 Model Selection

In order for the Bayesian framework to be useful in an application, one needs to make a number of modeling decisions. First of all, one needs to choose the covariance function and likelihood function to be used. Furthermore, each may depend on a number of hyperparameters whose values also need to be determined. Both problems can be taken care of by the same methods, and are therefore both termed *model selection*. Obtaining the most likely model given the data can be done by maximizing the evidence (or marginal likelihood) $p(\mathcal{D})$ [MacKay, 1995]. The Laplace

approximation of $\Psi(\boldsymbol{f})$, discussed in Section 3.1 allows the evidence to be computed as an explicit expression:

$$\begin{aligned} p(\mathcal{D}|\theta) &= \int p(\mathcal{D}|\boldsymbol{f})p(\boldsymbol{f})d\boldsymbol{f} = \int \exp \Psi(\boldsymbol{f}) \\ &\approx \int \exp\left[\Psi(\hat{\mathbf{f}}) - \frac{1}{2}(\boldsymbol{f}-\hat{\mathbf{f}})^T[K^{-1}+W](\boldsymbol{f}-\hat{\mathbf{f}})\right] \\ &= \exp(\Psi(\hat{\mathbf{f}})) \int \exp\left[-\frac{1}{2}(\boldsymbol{f}-\hat{\mathbf{f}})^T[K^{-1}+W](\boldsymbol{f}-\hat{\mathbf{f}})\right] \\ &= \exp(\Psi(\hat{\mathbf{f}}))(2\pi)^{n/2}|(K^{-1}+W)^{-1}| \end{aligned} \tag{20}$$

As the logarithm is a monotonic function, maximizing the evidence is equivalent to minimizing the negative log evidence, which is approximated by taking the logarithm of Eq. (20) resulting in

$$-\ln p(\mathcal{D}|\theta) = \frac{1}{2}\hat{\mathbf{f}}^T K^{-1}\hat{\mathbf{f}} - \ln p(\mathcal{D}|\hat{\mathbf{f}}) + \frac{1}{2}\ln(|\mathbf{I}+KW|) \tag{21}$$

Several methods can be used to minimize the negative log evidence in Eq. (21). One could do a grid search, but this becomes expensive with a large number of hyperparameters. Here, we consider a gradient based method which allows one to optimize a large number of parameters. The derivatives together with the evaluation of the evidence can be given to a gradient based minimizer for finding the optimal values of the hyperparameters. Note, however, that the hyperparameters κ, σ have the additional constraint that their values should not be negative. By taking instead the logarithm of these parameters, i.e., taking the set $\{\ln\kappa, \ln\sigma, b\}$ as the variables to tune, we convert the constrained optimization problem in an unconstrained optimization problem. The derivatives of Eq. (21) to the hyperparameters $\{\ln\kappa, \ln\sigma, b\}$ are given in Appendix D.

5 Empirical Results

5.1 Methodology

5.1.1 Data set

We use data from Arehart et al. [2007], who collected pairwise preference data using listener experiments. In this study, participants include 14 subjects with normal-hearing and 18 subjects with hearing loss of presumed cochlear origin. The stimuli presented were two sets (one male, one female talker) of concatenated sentences from the hearing-in-noise-test (HINT) [Nilson et al., 1994]. The sentences were subjected to three types of degradation: symmetric peak-clipping, symmetric center-clipping, and additive stationary speech-shaped noise. The clipping conditions were included as they are related to distortion mechanisms found in hearing aids. Peak clipping is related to arithmetic, amplifier, and transducer saturation. Center clipping is related to numeric underflow and to the effects of noise-suppression signal processing in reducing the intensity of low-level signal components. Each stimuli was subjected to 24 distortion conditions, i.e., to 8 levels of each type of degradation.

Each subject participated in 3 one-hour sessions. During each session three blocks of 72 paired comparisons were presented, of which the first block in the first session was a trial block. Hence, in total 576 paired comparisons were collected for each participant.

In our approach to model and predict speech quality for arbitrary degradation mechanisms we need to make assumptions about what factors are dominant in forming quality judgments. Here, we follow Arehart et al. [2007] and assume that audibility may be an important factor in the perceptual judgment of quality by subjects. The coherence speech intelligibility index (CSII) approach of Kates and Arehart [2005] gives a procedure to take into account audibility factors and computes for each sound sample three features, namely CSII_{Low}, CSII_{Mid}, and $\text{CSII}_{\text{High}}$. We use the CSII approach to represent the pairwise preference data of Arehart et al. [2007], i.e., the data consists of pairwise experiments (x_1, x_2, d) of two sound samples x_1, x_2 and subject decision

$d \in \{1, -1\}$ denoting whether $x_1 \succ x_2$ or $x_2 \succ x_1$, respectively. Each sound sample is represented by three features CSII_{Low}, CSII_{Mid}, and $\text{CSII}_{\text{High}}$, which can take values in $[0, 1]^3$ although they are not uniformly distributed.

5.1.2 K-fold cross-validation

In K-fold cross-validation [Kohavi, 1995], a data set $\mathcal{D}$ is partitioned into K mutually exclusive subsets $\mathcal{D}_1, \ldots, \mathcal{D}_K$ of approximately equal size. The classifier is trained K times (the number of folds), each time $t \in \{1, \ldots, K\}$ it is trained on the data set $\mathcal{D} \setminus \mathcal{D}_t$ and validated on the data set D_t. Each sample is thus used for training and used exactly once for testing. The K results from the folds can be combined to produce a single estimate. For example, the cross-validation estimate of accuracy is the overall number of correct classifications, divided by the number of instances $|\mathcal{D}|$. Formally, let $\mathcal{C}(\mathcal{D}, x)$ be the classifier that returns a label for sample x when trained on $\mathcal{D}$. Let $\mathcal{D}_{(i)}$ be the test set that contains instance $(v_{i,1}, v_{i,2}, d_i)$, then

$$acc_{CV} = \frac{1}{|\mathcal{D}|} \sum_{(v_{i,1}, v_{i,2}, d_i) \in \mathcal{D}} \delta(\mathcal{C}(\mathcal{D} \setminus \mathcal{D}_{(i)}), d_i)$$

where $\delta(i, j) = 1$ iff $i = j$ and $\delta(i, j) = 0$ iff $i \neq j$. The prediction error (PE) is $1 - acc_{CV}$. 10-fold cross-validation is a common choice and is also used here.

In order to determine whether one classifier $\hat{f}_A$ significantly improves upon another classifier $\hat{f}_B$ when training with K-fold cross-validation, one can use McNemar's test [Dietterich, 1998, Everitt, 1977, Salzberg, 1997]. McNemar's test only focuses on the outcomes of the classifiers that are different, i.e., outcomes that are classified both correctly or incorrectly by both classifiers are disregarded. In our case, many comparisons are very easy to classify, e.g., no noise versus a lot of peak clipping, and will be classified correctly by any reasonable classifier and will be disregarded with McNemar's test. McNemar's test focusses on the hard to classify cases.

K-fold cross-validation results in a classification for each sample, as each sample is used exactly once for testing, from which we can construct the following contingency table (cf. [Dietterich, 1998]):

Number of examples misclassified by both $\hat{f}_A$ and $\hat{f}_B$	Number of examples misclassified by $\hat{f}_A$ but not by $\hat{f}_B$
Number of examples misclassified by $\hat{f}_B$ but not by $\hat{f}_A$	Number of examples misclassified by neither $\hat{f}_A$ nor $\hat{f}_B$.

We will use the notation

n_{00}	n_{01}
n_{10}	n_{11}

where $n = n_{00} + n_{01} + n_{10} + n_{11}$ is the total number of samples.

Under the null hypothesis that both algorithms perform equally well, the two algorithms should have the same error rate, which means that $n_{01} = n_{10}$. McNemar's test is based on a χ^2 test for goodness of fit that compares the distribution of counts expected under the null hypothesis to the observed counts. The expected counts under the null hypothesis are

n_{00}	$(n_{01} + n_{10})/2$
$(n_{01} + n_{10})/2$	n_{11}

The following statistic is distributed (approximately) as χ^2 with 1 degree of freedom; it incorporates a "continuity correction" term (of -1 in the numerator) to account for the fact that the statistic is discrete while the χ^2 distribution is continuous:

$$\frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \tag{22}$$

If the null hypothesis is correct, then the probability that this quantity is greater than $\chi^2_{1,0.95} = 3.841459$ is less than 0.05. Then we may reject the null hypothesis in favor of the hypothesis that the two algorithms have different performance.

5.2 Results

In this section we compare three classifiers for the pairwise comparison data of Arehart et al. [2007]. The first classifier is the Q_3 metric reported in Arehart et al. [2007], which is a minimum mean-squared error fit to the normal-hearing subjects' quality ratings and is given by

$$Q_3 = \frac{1}{1+e^{-c}} \text{ with } c = -4.56 + 2.41 \cdot \text{CSII}_{\text{Low}} + 2.16 \cdot \text{CSII}_{\text{Mid}} + 1.73 \cdot \text{CSII}_{\text{High}} \tag{23}$$

The same Q_3 metric was used for the normal-hearing and the hearing-impaired subjects. The metric is therefore based on the assumptions that the same low, mid, and high-level weights are appropriate for both subject populations, and that the audiogram embedded in the SII calculations is sufficient to explain the group differences. Note, that the Q_3 measure does not incorporate a response bias and that the constant -4.56 is irrelevant as two Q_3 values are subtracted when determining the preference between two samples.

Using the Bayesian framework discussed in Section 2 we trained two other classifiers. A Gaussian Process with a linear kernel (LK) and a Gaussian Process with a Gaussian kernel (GK) (cf. Appendix B). The linear kernel corresponds to probit regression for each individual participant. Both kernels were trained using 10-fold cross-validation and were compared to each other using the McNemar test (cf. Section 5.1.2). The results of the comparisons are shown in Table 1.

The first column is an identifier indicating the participant. The top half (from 'nh1' to 'nh14') are the 14 normal-hearing participants. The bottom half (from 'hi1' to 'hi18') are the 18 hearing-impaired participants. The second column reports the percentage of biased pairs of experiments, i.e., with $x_i \neq x_j$ the percentage of pairs such that (x_i, x_j, d) and (x_j, x_i, d) holds for $d \in \{1, -1\}$. The third column reports a consistency check *independent from the response bias*. For this, we counted for all quadruples $A, B, C, D \in X$ distinct whether the subject's preferences $d_i \in \{-1, 1\}$ in $(A, B, d_1), (C, B, d_2), (A, D, d_3), (C, D, d_4)$ are consistent with some total ordering over (A, B, C, D).[1] Columns 4, 5, and 6 report the prediction error for the three classifiers on each data set corresponding to a single participant. The last three columns report the comparison results between the classifiers using McNemar's test. Here, 'p' is the reported p-value, 's' is the number of successes (i.e., the number of experiments correctly predicted by the first named classifier, but wrongly predicted by the second named classifier), and 'f' the number of failures (i.e., the number of experiments wrongly predicted by the first named classifier, but correctly predicted by the second named classifier). Finally, the results for all normal-hearing and the results for all hearing-impaired subjects are collected and shown in the rows 'pool', i.e., average percentages and prediction errors, and pooled McNemar test results.

The second column shows that normal-hearing subjects have a lower response bias than hearing-impaired subjects (mean of 17.4% versus 25.9%) and there is less variability within the group (standard deviation 0.95% versus 1.69%). The results are not shown here, but incorporating a response bias significantly improved the model for hearing-impaired subjects, i.e., 7 cases for a GK model with bias versus a GK model without bias. Analogously, normal-hearing subjects are more consistent in their responses (mean 0.41% versus 0.91% with standard deviations 0.08% and 0.18% respectively). Note, that the high percentage in consistency comes from the large number of obvious preference relations and a high baseline of 12.5% for a random guesser.

Looking at the results for the normal-hearing subjects, we see that the classification performances of the three classifiers is very similar. Sometimes one classifier performs better than another, sometimes worse. The classifier performance of one classifier, however, is never significantly better than another classifier as all reported p-values are greater or equal than 0.05. This changes, however, when we consider the classification performance on the subdata corresponding

[1] Note, only $(d_1, d_2, d_3, d_4) \in \{(-1, 1, 1, -1), (1, -1, -1, 1)\}$ does not lead to a total order over (A, B, C, D).

Table 1: Prediction error (PE) of several classifiers on the normal-hearing and hearing-impaired data and comparisons using McNemar's test with 'p' the p-value, 's' the successes, and 'f' the failures of the first method versus the second method.

Subj.	**%**	**%**	**Q3**	**LK**	**GK**	**Q3 vs LK**			**Q3 vs GK**			**LK vs GK**		
name	**bias**	**incons.**	**PE**	**PE**	**PE**	**p**	**s**	**f**	**p**	**s**	**f**	**p**	**s**	**f**
nh1	21.7	0.11	0.15	0.14	0.14	0.81	8	10	0.68	13	10	0.23	8	3
nh2	14.5	0.21	0.13	0.11	0.11	0.14	14	24	0.07	15	28	0.58	5	8
nh3	13.4	0.27	0.12	0.12	0.12	0.86	16	14	0.88	22	20	0.86	16	16
nh4	10.1	0.14	0.12	0.12	0.12	1.00	20	19	0.55	25	20	0.39	8	4
nh5	15.9	0.21	0.14	0.15	0.15	0.15	16	8	0.86	15	15	0.17	9	17
nh6	17.4	0.20	0.14	0.14	0.14	0.86	14	16	0.57	22	27	0.74	16	19
nh7	18.5	0.61	0.14	0.18	0.18	1.00	27	28	0.42	34	42	0.42	24	31
nh8	22.8	1.28	0.21	0.21	0.21	0.88	22	22	0.19	24	35	0.14	17	28
nh9	14.1	0.32	0.16	0.16	0.16	0.87	19	21	1.00	19	18	0.78	26	23
nh10	15.9	0.33	0.14	0.14	0.14	1.00	11	10	0.58	13	17	0.40	9	14
nh11	18.1	0.68	0.16	0.18	0.18	0.05	22	10	0.17	27	17	0.84	11	13
nh12	19.9	0.74	0.15	0.15	0.15	1.00	19	18	0.20	25	36	0.10	17	29
nh13	22.5	0.27	0.15	0.15	0.15	0.87	19	19	1.00	20	19	1.00	5	4
nh14	18.5	0.33	0.14	0.12	0.12	0.16	12	21	0.26	15	23	1.00	6	5
pool	17.4	0.41	0.15	0.15	0.15	1.00	239	240	0.14	289	327	0.07	177	214
hi1	37.7	0.82	0.26	0.19	0.19	0.00	56	94	0.00	29	115	0.00	17	65
hi2	21.7	0.57	0.16	0.11	0.11	0.01	22	46	0.00	20	45	1.00	13	14
hi3	18.8	0.82	0.18	0.16	0.16	0.27	28	38	0.28	37	48	1.00	24	25
hi4	22.8	0.43	0.18	0.15	0.15	0.03	30	51	0.00	24	52	0.32	15	22
hi5	32.2	0.84	0.20	0.14	0.14	0.00	33	64	0.00	33	64	0.80	8	8
hi6	27.9	1.43	0.24	0.23	0.23	0.34	31	40	0.06	37	56	0.22	22	32
hi7	18.1	0.44	0.14	0.11	0.11	0.04	20	36	0.02	17	35	0.79	6	8
hi8	21.7	1.23	0.18	0.17	0.17	0.88	22	24	0.11	22	35	0.14	17	28
hi9	18.1	0.59	0.15	0.12	0.12	0.03	15	31	0.01	15	35	0.62	16	20
hi10	15.6	0.26	0.14	0.12	0.12	0.01	7	21	0.01	12	31	0.42	10	15
hi11	29.3	0.22	0.16	0.12	0.12	0.01	22	44	0.00	18	53	0.03	9	22
hi12	35.5	0.31	0.19	0.14	0.14	0.00	28	56	0.00	18	79	0.00	15	48
hi13	35.9	2.00	0.24	0.22	0.22	0.27	44	56	0.00	45	78	0.00	15	36
hi14	25.0	0.66	0.18	0.13	0.13	0.00	21	52	0.00	16	58	0.03	6	17
hi15	14.1	0.22	0.13	0.14	0.14	0.80	32	29	0.72	33	37	0.19	7	14
hi16	30.1	2.05	0.31	0.23	0.23	0.00	63	109	0.00	45	108	0.06	27	44
hi17	31.9	3.04	0.24	0.22	0.22	0.28	37	48	0.19	36	49	0.75	4	6
hi18	30.4	0.49	0.18	0.15	0.15	0.05	19	34	0.04	24	42	0.69	11	14
pool	25.9	0.91	0.19	0.16	0.16	0.00	530	873	0.00	481	1020	0.00	242	438

to the hearing-impaired participants. The Q_3 metric is significantly outperformed by the LK and GK classifiers on 11 and 13 participants respectively. The LK classifier is again significantly outperformed by the GK classifier on 5 participants. Note that the Q_3 metric uses a linear model fitted to the group of normal-hearing subjects for the group of hearing-impaired subjects, whereas the GK and LK models are fitted for each subject individually.

It follows from these results that the prediction of speech quality for hearing-impaired subjects and arbitrary degradation mechanisms can be significantly improved by (1) personalization (i.e., using individual preferences as well as modeling a response bias significantly improved the model), and by (2) allowing nonlinear relationships in the model. For normal hearing subjects simple logistic regression techniques can be used as no significant improvements could be obtained when

using a more complex model. This demonstrates that there are significant differences between the groups of normal-hearing and hearing-impaired subjects and one should be careful generalizing models learned from/fitted on normal-hearing subjects to hearing-impaired subjects.

5.3 Visualization

To demonstrate the nonlinear perception in hearing-impaired subjects we show in Figure 1 the elicitated utility function of one of the subjects with a significant improvement in prediction quality. As the CSII_{Mid} features were found to be highly correlated with the CSII_{Low} and $\text{CSII}_{\text{High}}$ features we used the following linear regression relation (fitted for subject 'hi12')

$$\text{CSII}_{\text{Mid}}{}' = -0.0134 + 0.6555 \cdot \text{CSII}_{\text{Low}} + 0.5351 \cdot \text{CSII}_{\text{High}} \tag{24}$$

to effectively reduce our graph to 3-dimensions. Figure 1 shows the hyperplane in terms of CSII_{Low}, $\text{CSII}_{\text{High}}$, and CSII_{Mid}' features as given by Eq. (24), the utility function on this hyperplane, and the samples projected on the contour plot of the utility function.

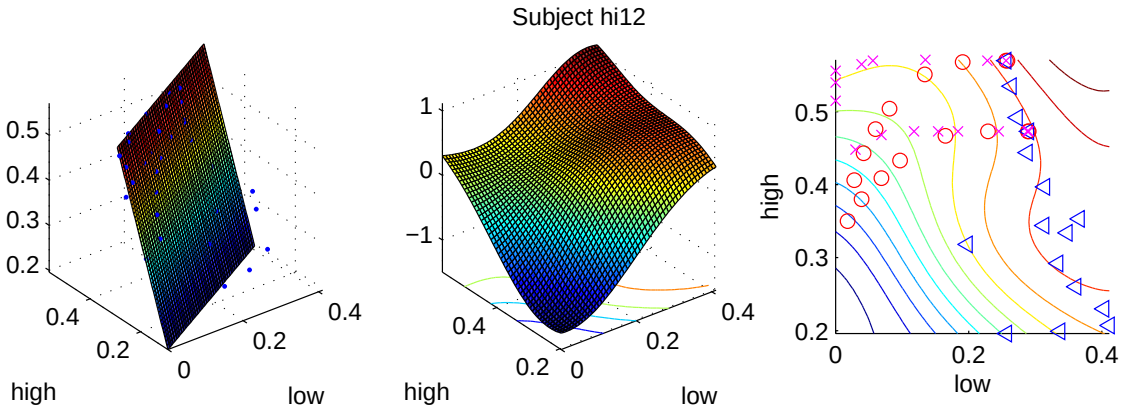

Figure 1: Utility function elicitated for subject 'hi12'. Left: hyperplane formed by linear regression of CSII_{Mid} in terms of CSII_{Low} and $\text{CSII}_{\text{High}}$ features. Middle: Utility function on hyperplane. Right: contour plot with sound samples distorted with noise (circle), peak clipping (triangle), and center clipping (cross).

5.4 Validation of Coherence

One of the questions that may rise from the results reported in Table 1 and Figure 1 is that the nonlinear behaviour in perception of hearing-impaired subjects is a result of the use of the coherence measure that transforms the sound features based on the audiogram of the subject. In order to validate the coherence based approach we compared a Gaussian Process having a Gaussian kernel on the hearing-impaired subjects' data using (1) the sound features present in the data set (i.e., the coherence measure incorporating the audiogram), with (2) the sound features following from the coherence measure without incorporating the audiogram (i.e., the same features used for normal-hearing subjects). The comparison results are shown in Table 2, which shows that incorporating the audiogram in the coherence measure significantly improved results for 3 hearing-impaired subjects. In some way, this validates the approach followed of using the coherence measure incorporating the audiogram of the subject.

Table 2: Prediction error (PE) of a Gaussian Process with Gaussian kernel for hearing-impaired using coherence based featues (GK) and features used by normal-hearing subjects (GK-map). Comparisons are made using McNemar's test with 'p' the p-value, 's' the successes, and 'f' the failures of the first method versus the second method.

Subj. name	**GK PE**	**GK-map PE**	**GK vs GK-map**		
			p	**s**	**f**
hi1	0.11	0.13	0.04	27	13
hi2	0.11	0.12	0.65	11	8
hi3	0.16	0.16	0.71	13	16
hi4	0.14	0.12	0.11	2	8
hi5	0.14	0.15	0.58	8	5
hi6	0.21	0.21	1.00	7	6
hi7	0.11	0.11	0.62	3	1
hi8	0.15	0.16	0.38	13	8
hi9	0.12	0.14	0.07	25	13
hi10	0.11	0.11	1.00	12	11
hi11	0.10	0.11	0.54	14	10
hi12	0.09	0.14	0.00	40	11
hi13	0.18	0.21	0.08	30	17
hi14	0.11	0.12	0.61	19	15
hi15	0.12	0.13	0.71	34	30
hi16	0.20	0.21	0.28	40	30
hi17	0.22	0.21	0.71	30	34
hi18	0.15	0.17	0.04	30	15
pool	0.14	0.15	0.00	358	251

6 Conclusions

This study began with the premise that the perceived quality of sound is a central issue in the development of hearing aids and other communicating devices. Methods for correctly predicting the perceived quality of a subject would advance their development.

In this study we advocated a Bayesian framework using Gaussian Processes, which takes into account a response bias and inconsistencies in user preferences. We have demonstrated that predicting the perceived quality of a hearing-impaired subject can significantly be improved by (1) learning from the subject's own preferences, and (2) incorporating nonlinearities in perception in the model. No such improvements could be made for normal-hearing subjects, indicating significant differences between both groups of subjects.

Gaussian Processes have received increased attention in the machine learning community over the past decade and have successfully been applied in numerous applications. In the current study, we have demonstrated a principled approach for dealing with nonlinearities in quality perception and random variability in the judgments of hearing-impaired subjects. Several modeling choices were made that allow for further extensions. First, the kernel function can be extended by incorporating properties of the auditory system. Second, the framework can be extended to a full Bayesian framework, i.e., priors over hyperparameters instead of maximum likelihood for model selection. Third, different likelihood functions for absolute ratings or polytomous choice models Andrich [1978], can be incorporated into the framework for investigating the best response scale when learning user preferences. Fourth, the framework can be extended to a hierarchical model Gelman et al. [2003] such that preferences from normal-hearing subjects can also be properly integrated (from a Bayesian viewpoint) into the utility elicitation process of a hearing-impaired subject. Fifth, other feature constructing methods than coherence can directly be incorporated into the framework, which can be combined with kernels for automatic feature selection.

Acknowledgment

We thank Adriana Birlutiu, Bert de Vries, and Iman Mossavat for earlier discussions. The current research was funded by STW project 07605 and NWO VICI grant 639.023.604.

A Notation

$\boldsymbol{f}$	latent function
$\hat{\mathbf{f}}$	maximum apostiori (MAP) estimate
$\mathcal{D}$	data set of pairwise preferences
$p(\boldsymbol{f})$	prior distribution
K	covariance matrix
W	$-\nabla\nabla \ln p(\mathcal{D}\|\boldsymbol{f})$
$p(\mathcal{D}\|\boldsymbol{f})$	likelihood function
$\Phi(z)$	cumulative Gaussian, $\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)dx$
θ	hyperparameters
κ	covariance hyperparameter
σ	likelihood noise hyperparameter
b	likelihood bias hyperparameter
$(v_{i,1}, v_{i,2}, d_i)$	i-th pairwise preference experiment
k	experiment index
m	total number of experiments
n	total number of distinct samples

B Kernel Functions

B.1 Gaussian Kernel

The Gaussian Kernel or Squared Exponential Kernel is defined in Chu and Ghahramani [2005] as

$$k(x,y) = \exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n}(x_i - y_i)^2\right) = \exp\left(-\frac{\kappa}{2} \parallel x - y \parallel^2\right) \tag{25}$$

and the derivative with respect to $\ln \kappa$ is

$$\frac{\partial k(x,y)}{\partial \ln \kappa} = \frac{\partial}{\partial \ln \kappa} \exp\left(-\frac{\kappa}{2} \parallel x - y \parallel^2\right) = \exp\left(-\frac{\kappa}{2} \parallel x - y \parallel^2\right) \cdot \left(-\frac{\kappa}{2} \parallel x - y \parallel^2\right) \tag{26}$$

B.2 Linear Kernel

The linear kernel with one hyperparameter is specified as follows

$$k(x,y) = \frac{1}{\kappa}\left(1 + \sum_{i=1}^{n}(x_i y_i)\right) \tag{27}$$

and the derivative with respect to $\ln \kappa$ is

$$\frac{\partial k(x,y)}{\partial \ln \kappa} = \frac{\partial}{\partial \ln \kappa}\frac{1}{\kappa}\left(1 + \sum_{i=1}^{n}(x_i y_i)\right) = \frac{-1}{\kappa}\left(1 + \sum_{i=1}^{n}(x_i y_i)\right) \tag{28}$$

C Derivatives of the Likelihood Function $\ln p(\mathcal{D}|\boldsymbol{f})$

To obtain the optimal hyperparemeters in the Bayesian framework considered using a gradient based optimization method, we will need derivatives of the log likelihood function $\ln p(\mathcal{D}|\boldsymbol{f})$ with respect to the function $\{f(x_i)\}$ and the likelihood parameters $\{\ln \sigma, b\}$.

C.1 Derivative of $\ln p(\mathcal{D}|\boldsymbol{f})$ with respect to the function values $f(x_i)$

First, we consider the individual observations, then use these results to compute the derivative of the log likelihood function.

C.1.1 The loss function

The quantity $-\ln p(v_{k,1}, v_{k,2}, d_k | f(v_{k,1}), f(v_{k,2}))$ is usually referred to as the loss function, i.e., $-\ln \Phi(z_k)$ and denoted $\ell(v_{k,1}, v_{k,2}, d_k, f(v_{k,1}), f(v_{k,2}))$. The first order derivative of the loss function is as follows

$$\begin{aligned}\frac{\partial -\ln \Phi(z_k)}{\partial f(x_i)} &= \frac{\partial -\ln \Phi(z_k)}{\partial z_k}\frac{\partial z_k}{\partial f(x_i)} = \frac{\partial -\ln \Phi(z_k)}{\partial z_k}\frac{d_k s_k(x_i)}{\sqrt{2}\sigma} \\ &= \frac{-\Phi(z_k)'}{\Phi(z_k)}\frac{d_k s_k(x_i)}{\sqrt{2}\sigma} = \frac{-\mathcal{N}(z_k;0,1)}{\Phi(z_k)}\frac{d_k s_k(x_i)}{\sqrt{2}\sigma}\end{aligned} \tag{29}$$

where we used the chain rule $f(g(x))' = f'(g(x)) \cdot g'(x)$ and $s_k(x_i) = 1$ iff x_i is the first sample in experiment k, $s_k(x_i) = -1$ iff x_i is the second sample in experiment k, and $s_k(x_i) = 0$ otherwise. The second order derivative of the loss function is as follows

$$\begin{aligned}\frac{\partial^2 -\ln \Phi(z_k)}{\partial f(x_i)\partial f(x_j)} &= \frac{\partial}{\partial f(x_j)}\left(\frac{\partial -\ln \Phi(z_k)}{\partial f(x_i)}\right) \\ &= \frac{\partial}{\partial f(x_j)}\left(\frac{-\mathcal{N}(z_k;0,1)}{\Phi(z_k)}\frac{d_k s_k(x_i)}{\sqrt{2}\sigma}\right) \\ &= \frac{\partial}{\partial z_k}\left(\frac{-\mathcal{N}(z_k;0,1)}{\Phi(z_k)}\frac{d_k s_k(x_i)}{\sqrt{2}\sigma}\right)\frac{\partial z_k}{\partial\ f(x_j)} \\ &= \frac{\partial}{\partial z_k}\left(\frac{-\mathcal{N}(z_k;0,1)}{\Phi(z_k)}\right)\frac{d_k s_k(x_i)}{\sqrt{2}\sigma}\frac{\partial z_k}{\partial f(x_j)} \\ &= \frac{\partial}{\partial z_k}\left(\frac{-\mathcal{N}(z_k;0,1)}{\Phi(z_k)}\right)\frac{d_k^2 s_k(x_i)}{\sqrt{2}\sigma}\frac{s_k(x_j)}{\sqrt{2}\sigma} \\ &= \frac{\partial}{\partial z_k}\left(\frac{-\mathcal{N}(z_k;0,1)}{\Phi(z_k)}\right)\frac{s_k(x_i)s_k(x_j)}{2\sigma^2} \\ &= \left(\frac{\mathcal{N}^2(z_k;0,1)}{\Phi^2(z_k)} + \frac{z_k\mathcal{N}(z_k;0,1)}{\Phi(z_k)}\right)\frac{s_k(x_i)s_k(x_j)}{2\sigma^2}\end{aligned} \tag{30}$$

where we used

$$\begin{aligned}\frac{\partial}{\partial z_k}\left(\frac{-\mathcal{N}(z_k;0,1)}{\Phi(z_k)}\right) &= (-\mathcal{N}(z_k;0,1))' \cdot \Phi^{-1}(z_k) + (-\mathcal{N}(z_k;0,1)) \cdot (\Phi^{-1}(z_k))' \\ &= \frac{z_k\mathcal{N}(z_k;0,1)}{\Phi(z_k)} + \frac{\mathcal{N}^2(z_k;0,1)}{\Phi^2(z_k)}\end{aligned} \tag{31}$$

Using the same techniques we find that

$$\begin{aligned}\frac{\partial^3 -\ln \Phi(z_k)}{\partial f(x_h)\partial f(x_i)\partial f(x_j)} = &\frac{-d_k s_k(x_h)s_k(x_i)s_k(x_j)}{(\sqrt{2}\sigma)^3}\cdot \\ &\left(\frac{2\mathcal{N}^3(z_k;0,1)}{\Phi^3(z_k)} + \frac{3z_k\mathcal{N}^2(z_k;0,1)}{\Phi^2(z_k)} + \frac{(z_k^2-1)\mathcal{N}(z_k;0,1)}{\Phi(z_k)}\right)\end{aligned} \tag{32}$$

C.1.2 The Log Likelihood

$$\begin{aligned}\frac{\partial \ln p(\mathcal{D}|\boldsymbol{f})}{\partial f(x_i)} &= \frac{\partial \ln \prod_{k=1}^m \Phi(z_k)}{\partial f(x_i)} = \frac{\partial \sum_{k=1}^m \ln \Phi(z_k)}{\partial f(x_i)} \\ &= \sum_{k=1}^m \frac{\partial \ln \Phi(z_k)}{\partial f(x_i)} = \sum_{k=1}^m \frac{d_k s_k(x_i)\mathcal{N}(z_k;0,1)}{\sqrt{2}\sigma\Phi(z_k)}\end{aligned} \tag{33}$$

$$\frac{\partial^2 \ln p(\mathcal{D}|\boldsymbol{f})}{\partial f(x_i)\partial f(x_j)} = \sum_{k=1}^{m} \frac{\partial^2 \ln \Phi(z_k)}{\partial f(x_i)\partial f(x_j)} = \sum_{k=1}^{m} \frac{-s_k(x_i)s_k(x_j)}{2\sigma^2}\left(\frac{\mathcal{N}^2(z_k;0,1)}{\Phi^2(z_k)} + \frac{z_k\mathcal{N}(z_k;0,1)}{\Phi(z_k)}\right) \tag{34}$$

$$\begin{aligned}\frac{\partial^3 \ln p(\mathcal{D}|\boldsymbol{f})}{\partial f(x_h)\partial f(x_i)\partial f(x_j)} = \sum_{k=1}^{m} \frac{d_k s_k(x_h)s_k(x_i)s_k(x_j)}{(\sqrt{2}\sigma)^3}\cdot \\ \left(\frac{2\mathcal{N}^3(z_k;0,1)}{\Phi^3(z_k)} + \frac{3z_k\mathcal{N}^2(z_k;0,1)}{\Phi^2(z_k)} + \frac{(z_k^2-1)\mathcal{N}(z_k;0,1)}{\Phi(z_k)}\right)\end{aligned} \tag{35}$$

Note that when $x_i \neq x_j$, the third partial derivative is 0, when $x_h \neq x_i \wedge x_h \neq x_j$.

C.2 Derivative of $\ln p(\mathcal{D}|\boldsymbol{f})$ with respect to the hyperparameters $\ln \sigma, b$

First, we compute the derivative of $\ln p(\mathcal{D}|\boldsymbol{f})$ with respect to the noise parameter $\ln \sigma$. We use $\nabla \ln p(\mathcal{D}|\boldsymbol{f})$ and $\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{f})$to denote the first and second derivative of $\ln p(\mathcal{D}|\boldsymbol{f})$ with respect to the values $\{f(x_i)\}$ given in Eqs. (33) and (34).

$$\begin{aligned}\frac{\partial \ln p(\mathcal{D}|\boldsymbol{f})}{\partial \ln \sigma} &= \frac{\partial \sum_{k=1}^{m} \ln \Phi(z_k)}{\partial \ln \sigma} = \sum_{k=1}^{m} \frac{\partial \ln \Phi(z_k)}{\partial \ln \sigma} \\ &= \sum_{k=1}^{m} \frac{\partial \ln \Phi(z_k)}{\partial z_k}\frac{\partial z_k}{\partial \ln \sigma} = \sum_{k=1}^{m} \frac{-z_k\mathcal{N}(z_k;0,1)}{\Phi(z_k)}\end{aligned} \tag{36}$$

$$\begin{aligned}\frac{\partial \nabla \ln p(\mathcal{D}|\boldsymbol{f})}{\partial \ln \sigma} &= \frac{\partial \nabla \ln p(\mathcal{D}|\boldsymbol{f})}{\partial \ln \sigma}\Big|_{explicit} + \sum_{k=1}^{m} \frac{\partial \nabla \ln p(\mathcal{D}|\boldsymbol{f})}{\partial z_k}\frac{\partial z_k}{\partial \ln \sigma} \\ &= -\nabla \ln p(\mathcal{D}|\boldsymbol{f}) - \sum_{k=1}^{m} z_k \frac{\partial \nabla \ln p(\mathcal{D}|\boldsymbol{f})}{\partial z_k}\end{aligned} \tag{37}$$

$$\begin{aligned}\frac{\partial \nabla\nabla \ln p(\mathcal{D}|\boldsymbol{f})}{\partial \ln \sigma} &= \frac{\partial \nabla\nabla \ln p(\mathcal{D}|\boldsymbol{f})}{\partial \ln \sigma}\Big|_{explicit} + \sum_{k=1}^{m} \frac{\partial \nabla\nabla \ln p(\mathcal{D}|\boldsymbol{f})}{\partial z_k}\frac{\partial z_k}{\partial \ln \sigma} \\ &= -2\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{f}) - \sum_{k=1}^{m} z_k \frac{\partial \nabla\nabla \ln p(\mathcal{D}|\boldsymbol{f})}{\partial z_k}\end{aligned} \tag{38}$$

where we used

$$\begin{aligned}\frac{\partial}{\partial \ln \sigma}\frac{\partial \ln p(\mathcal{D}|\boldsymbol{f})}{\partial f(x_i)}\Big|_{explicit} &= \frac{\partial}{\partial \ln \sigma}\sum_{k=1}^{m} \frac{s_k(x_i)\mathcal{N}(z_k;0,1)}{\sqrt{2}\sigma\Phi(z_k)} = \frac{\partial}{\partial z}\sum_{k=1}^{m} \frac{s_k(x_i)\mathcal{N}(z_k;0,1)}{\sqrt{2}\Phi(z_k)}\exp(-z) \\ &= -\sum_{k=1}^{m} \frac{s_k(x_i)\mathcal{N}(z_k;0,1)}{\sqrt{2}\Phi(z_k)}\exp(-z) = -\sum_{k=1}^{m} \frac{s_k(x_i)\mathcal{N}(z_k;0,1)}{\sqrt{2}\sigma\Phi(z_k)} \\ &= -\frac{\partial \ln p(\mathcal{D}|\boldsymbol{f})}{\partial f(x_i)}\end{aligned} \tag{39}$$

$$\begin{aligned}\frac{\partial}{\partial z_k}\frac{\partial \ln p(\mathcal{D}|\boldsymbol{f})}{\partial f(x_i)} &= \frac{\partial}{\partial z_k}\sum_{k'=1}^{m} \frac{s_{k'}(x_i)}{\sqrt{2}\sigma}\frac{\mathcal{N}(z_{k'};0,1)}{\Phi(z_{k'})} = \sum_{k'=1}^{m} \frac{\partial}{\partial z_k}\frac{s_{k'}(x_i)}{\sqrt{2}\sigma}\frac{\mathcal{N}(z_{k'};0,1)}{\Phi(z_{k'})} \\ &= \frac{\partial}{\partial z_k}\frac{s_k(x_i)}{\sqrt{2}\sigma}\frac{\mathcal{N}(z_k;0,1)}{\Phi(z_k)} = \frac{-s_k(x_i)}{\sqrt{2}\sigma}\left(\frac{\mathcal{N}^2(z_k;0,1)}{\Phi^2(z_k)} + \frac{z_k\mathcal{N}(z_k;0,1)}{\Phi(z_k)}\right)\end{aligned} \tag{40}$$

$$\frac{\partial}{\partial z_k}\frac{\partial^2 \ln p(\mathcal{D}|\boldsymbol{f})}{\partial f(x_i)\partial f(x_j)} = \frac{s_k(x_i)s_k(x_j)}{2\sigma^2}\cdot \left(2\frac{\mathcal{N}^3(z_k;0,1)}{\Phi^3(z_k)} + 3z_k\frac{\mathcal{N}^2(z_k;0,1)}{\Phi^2(z_k)} + \frac{(z_k^2-1)\mathcal{N}(z_k;0,1)}{\Phi(z_k)}\right) \tag{41}$$

Next, we compute the derivative of the likelihood function $\ln p(\mathcal{D}|\boldsymbol{f})$ with respect to the bias hyperparameter b.

$$\frac{\partial \ln p(\mathcal{D}|\boldsymbol{f})}{\partial b} = \sum_{k=1}^{m}\frac{\partial \ln \Phi(z_k)}{\partial z_k}\frac{\partial z_k}{\partial b} = \sum_{k=1}^{m}\frac{\mathcal{N}(z_k;0,1)}{\Phi(z_k)}\cdot\frac{-d_k}{\sqrt{2}\sigma} \tag{42}$$

$$\frac{\partial \nabla \ln p(\mathcal{D}|\boldsymbol{f})}{\partial b} = 0 + \sum_{k=1}^{m}\frac{\partial \nabla \ln p(\mathcal{D}|\boldsymbol{f})}{\partial z_k}\frac{\partial z_k}{\partial b} = \sum_{k=1}^{m}\frac{\partial \nabla \ln p(\mathcal{D}|\boldsymbol{f})}{\partial z_k}\frac{-d_k}{\sqrt{2}\sigma} \tag{43}$$

$$\frac{\partial \nabla\nabla \ln p(\mathcal{D}|\boldsymbol{f})}{\partial b} = 0 + \sum_{k=1}^{m}\frac{\partial \nabla\nabla \ln p(\mathcal{D}|\boldsymbol{f})}{\partial z_k}\frac{\partial z_k}{\partial b} = \sum_{k=1}^{m}\frac{\partial \nabla\nabla \ln p(\mathcal{D}|\boldsymbol{f})}{\partial z_k}\frac{-d_k}{\sqrt{2}\sigma} \tag{44}$$

where the last two equations can be rewritten using the results of Eqs. (40) and (41).

D Derivatives of the Negative Log Evidence $-\ln p(\mathcal{D}|\theta)$

The negative log evidence is a function of the hyperparameters θ, but also of $\hat{\mathbf{f}}$, which depends on θ. Hence, when taking the derivative with respect to the log hyperparameters we need to sum the explicit and implicit partial derivatives:

$$\frac{\partial -\ln p(\mathcal{D}|\theta)}{\partial \ln \theta} = \frac{\partial -\ln p(\mathcal{D}|\theta)}{\partial \ln \theta}\Big|_{explicit} + \sum_{i=1}^{n}\frac{\partial -\ln p(\mathcal{D}|\theta)}{\partial \hat{\mathbf{f}}(x_i)}\cdot\frac{\partial \hat{\mathbf{f}}(x_i)}{\partial \ln \theta} \tag{45}$$

D.1 Derivative of $-\ln p(\mathcal{D}|\theta)$ with respect to $\ln \kappa$

The explicit part of the derivative with respect to the covariance matrix parameter $\ln \kappa$.

$$\frac{\partial}{\partial \ln \kappa}\frac{1}{2}\hat{\mathbf{f}}^T K^{-1}\hat{\mathbf{f}} = -\frac{1}{2}\hat{\mathbf{f}}^T K^{-1}\frac{\partial K}{\partial \ln \kappa}K^{-1}\hat{\mathbf{f}} \tag{46}$$

$$\begin{aligned}\frac{\partial}{\partial \ln \kappa}\frac{1}{2}\ln(|I+KW|) &= \frac{1}{2}\mathrm{tr}((I+KW)^{-1}\frac{\partial I+KW}{\partial \ln \kappa}) \\ &= \frac{1}{2}\mathrm{tr}((I+KW)^{-1}\frac{\partial K}{\partial \ln \kappa}W) \\ &= \frac{1}{2}\mathrm{tr}(W(I+KW)^{-1}\frac{\partial K}{\partial \ln \kappa}) \\ &= \frac{1}{2}\mathrm{tr}(((I+KW)W^{-1})^{-1}\frac{\partial K}{\partial \ln \kappa}) \\ &= \frac{1}{2}\mathrm{tr}(((W^{-1}+K)^{-1}\frac{\partial K}{\partial \ln \kappa})\end{aligned} \tag{47}$$

Next, the implicit part. First note that because $\hat{\mathbf{f}}$ is the maximum of the posterior we have

$$\frac{\partial}{\partial \hat{\mathbf{f}}(x_i)}\frac{1}{2}\hat{\mathbf{f}}^T K^{-1}\hat{\mathbf{f}} - \ln p(\mathcal{D}|\hat{\mathbf{f}}) = \frac{\partial \Psi(\hat{\mathbf{f}})}{\partial \hat{\mathbf{f}}(x_i)} = 0 \tag{48}$$

Therefore we can restrict to $\frac{1}{2}\ln(|I+KW|)$ for the implicit partial derivative.

$$\begin{aligned}\frac{\partial}{\partial\hat{\mathbf{f}}(x_i)}\frac{1}{2}\ln(|I+KW|) &= \frac{1}{2}\mathrm{tr}((I+KW)^{-1}\frac{\partial I+KW}{\partial\hat{\mathbf{f}}(x_i)})\\ &= \frac{1}{2}\mathrm{tr}((I+KW)^{-1}\frac{\partial I+KW}{\partial W}\cdot\frac{\partial W}{\partial\hat{\mathbf{f}}(x_i)})\\ &= \frac{1}{2}\mathrm{tr}((I+KW)^{-1}K\frac{\partial W}{\partial\hat{\mathbf{f}}(x_i)})\\ &= \frac{1}{2}\mathrm{tr}((K^{-1}(I+KW))^{-1}\frac{\partial W}{\partial\hat{\mathbf{f}}(x_i)})\\ &= \frac{1}{2}\mathrm{tr}((K^{-1}+W)^{-1}\frac{\partial W}{\partial\hat{\mathbf{f}}(x_i)})\end{aligned} \tag{49}$$

Furthermore, because $\hat{\mathbf{f}} = K\nabla\ln p(D|\hat{\mathbf{f}})$ (cf. (3.17) in Rasmussen and Williams [2006]) we have

$$\begin{aligned}\frac{\partial\hat{\mathbf{f}}}{\partial\ln\theta} &= \frac{\partial K\nabla\ln p(\mathcal{D}|\hat{\mathbf{f}})}{\partial\ln\theta}\\ &= \frac{\partial K}{\partial\ln\theta}\nabla\ln p(\mathcal{D}|\hat{\mathbf{f}}) + K\frac{\partial\nabla\ln p(\mathcal{D}|\hat{\mathbf{f}})}{\partial\ln\theta}\\ &= \frac{\partial K}{\partial\ln\theta}\nabla\ln p(\mathcal{D}|\hat{\mathbf{f}}) + K\frac{\partial\nabla\ln p(\mathcal{D}|\hat{\mathbf{f}})}{\partial\ln\theta}\Big|_{explicit} + K\frac{\partial\nabla\ln p(\mathcal{D}|\hat{\mathbf{f}})}{\partial\hat{\mathbf{f}}}\frac{\partial\hat{\mathbf{f}}}{\partial\ln\theta}\\ &= \frac{\partial K}{\partial\ln\theta}\nabla\ln p(\mathcal{D}|\hat{\mathbf{f}}) + K\frac{\partial\nabla\ln p(\mathcal{D}|\hat{\mathbf{f}})}{\partial\ln\theta}\Big|_{explicit} - KW\frac{\partial\hat{\mathbf{f}}}{\partial\ln\theta}\end{aligned} \tag{50}$$

Note that the middle term (explicite derivative) is 0, when $\theta = \kappa$. In that case, collecting all terms with $\frac{\partial\hat{\mathbf{f}}}{\partial\ln\theta}$ on the left hand side and multiplying both sides with $(I+KW)^{-1}$ results in

$$\frac{\partial\hat{\mathbf{f}}}{\partial\ln\kappa} = (I+KW)^{-1}\frac{\partial K}{\partial\ln\kappa}\nabla\ln p(\mathcal{D}|\hat{\mathbf{f}}) \tag{51}$$

D.2 Derivative of $-\ln p(\mathcal{D}|\theta)$ with respect to $\ln\sigma$ and b

D.2.1 Derivative with respect to $\ln\sigma$

Note that

$$\begin{aligned}\frac{\partial}{\partial\ln\sigma} - \ln p(\mathcal{D}|\theta) &= \frac{\partial}{\partial\ln\sigma} - \ln p(\mathcal{D}|\hat{\mathbf{f}}) + \frac{\partial}{\partial\ln\sigma}\frac{1}{2}\ln(|I+KW|)\\ &= \frac{\partial}{\partial\ln\sigma} - \ln p(\mathcal{D}|\hat{\mathbf{f}}) + \frac{1}{2}\mathrm{tr}((K^{-1}+W)^{-1}\frac{\partial W}{\partial\ln\sigma})\end{aligned} \tag{52}$$

where we used

$$\begin{aligned}\frac{\partial}{\partial\ln\sigma}\frac{1}{2}\ln(|I+KW|) &= \frac{1}{2}\mathrm{tr}\left((I+KW)^{-1}\frac{\partial I+KW}{\partial\ln\sigma}\right) = \frac{1}{2}\mathrm{tr}\left((I+KW)^{-1}\frac{\partial KW}{\partial\ln\sigma}\right)\\ &= \frac{1}{2}\mathrm{tr}\left((I+KW)^{-1}K\frac{\partial W}{\partial\ln\sigma}\right) = \frac{1}{2}\mathrm{tr}\left((K^{-1}(I+KW))^{-1}\frac{\partial W}{\partial\ln\sigma}\right)\\ &= \frac{1}{2}\mathrm{tr}\left((K^{-1}+W)^{-1}\frac{\partial W}{\partial\ln\sigma}\right)\end{aligned} \tag{53}$$

As W (implicitly) depends on z_k and $\hat{\mathbf{f}}$, the derivative of W w.r.t. $\ln\sigma$ is given by

$$\frac{\partial}{\partial\ln\sigma}W = \frac{\partial W}{\partial\ln\sigma}\Big|_{explicit} + \sum_{k=1}^{m}\frac{\partial W}{\partial z_k}\frac{\partial z_k}{\partial\ln\sigma} + \sum_{i=1}^{n}\frac{\partial W}{\partial\hat{\mathbf{f}}(x_i)}\frac{\partial\hat{\mathbf{f}}(x_i)}{\partial\ln\sigma} \tag{54}$$

From Equation (50) it follows that

$$\begin{aligned}\frac{\partial \hat{\mathbf{f}}}{\partial \ln \sigma} &= (I + KW)^{-1} K \frac{\partial \nabla \ln p(\mathcal{D}|\hat{\mathbf{f}})}{\partial \ln \sigma}\Big|_{explicit} \\ &= (K^{-1} + W)^{-1} \frac{\partial \nabla \ln p(\mathcal{D}|\hat{\mathbf{f}})}{\partial \ln \sigma}\Big|_{explicit}\end{aligned} \tag{55}$$

D.2.2 Derivative with respect to b

Note that b can be any number, both positive and negative. Hence, we take the derivative w.r.t. b, not $\ln b$, for which holds that

$$\frac{\partial}{\partial b} - \ln p(\mathcal{D}|\theta) = \frac{1}{b}\frac{\partial}{\partial \ln b} - \ln p(\mathcal{D}|\theta) \tag{56}$$

D.3 Stable Computations

Several of the matrices used when computing the hyperparameters can be rewritten to make computations more stable using the matrix inversion lemma (cf. A.9 in Rasmussen and Williams [2006]).

$$Z = (W^{-1} + K)^{-1} = W^{\frac{1}{2}} W^{-\frac{1}{2}} (W^{-1} + K)^{-1} W^{-\frac{1}{2}} W^{\frac{1}{2}} = W^{\frac{1}{2}} (I + W^{\frac{1}{2}} K W^{\frac{1}{2}})^{-1} W^{\frac{1}{2}} \tag{57}$$

$$B^{-1} = (I + KW)^{-1} = I - K(W^{-1} + K)^{-1} = I - KZ \tag{58}$$

$$(K^{-1} + W)^{-1} = (K^{-1} + W^{\frac{1}{2}} I W^{\frac{1}{2}})^{-1} = K - KW^{\frac{1}{2}} (I + W^{\frac{1}{2}} K W^{\frac{1}{2}})^{-1} W^{\frac{1}{2}} K = K - KZK \tag{59}$$

References

D. Andrich. A rating formulation for ordered response categories. *Psychometrika*, 43:561–573, 1978.

K. Arehart, J. Kates, C. Anderson, and L. Harvey Jr. Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.*, 122 (2):115–1164, August 2007.

R. Bradley and M. Terry. Rank analysis of incomplete block designs, I. the method of paired comparisons. *Biometrika*, 39:324–345, 1952.

W. Chu and Z. Ghahramani. Preference Learning with Gaussian Processes. In *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

E. Czerwinski, A. Voishvillo, S. Alexandrov, and A. Terkhov. Multitone testing of sound system components: Some results and conclusions, Part 1: History and theory. *J. Audio. Eng. Soc.*, 49:1011–1048, 2001a.

E. Czerwinski, A. Voishvillo, S. Alexandrov, and A. Terkhov. Multitone testing of sound system components: Some results and conclusions, Part 2: Modeling and application. *J. Audio. Eng. Soc.*, 49:1181–1192, 2001b.

T. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

O. Dyrlund. Coherence measurements in hearing instruments, using different broadband signals. *Scanc. Audiol.*, 21:73–78, 1992.

B. Everitt. *The analysis of contingency tables*. Chapman and Hall, London, 1977.

A. Gabrielsen, B. Schenkman, and B. Hagerman. The effects of different frequency responses on sound quality judgments and speech intelligibility. *J. Speech. Hear. Res.*, 31:166–177, 1998.

A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman & Hall / CRC, 2003.

J. Kates. A test suite for hearing aid evaluation. *J. Rehab. Res. Dev.*, 27:255–278, 1990.

J. Kates and K. Arehart. Coherence and the speech intelligibility index. *J. Acoust. Soc. Am.*, 117: 2224–2237, 2005.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1137–1143, San Mateo, 1995. Morgan Kaufmann.

H. Levitt, E. Cudahy, H. Hwang, E. Kennedy, and C. Link. Towards a general measure of distortion. *J. Rehab. Res. Dev.*, 24:283–292, 1987.

R. Luce. *Individual Choice Behaviours: A Theoretical Analysis*. J. Wiley, New York, 1959.

D. MacKay. Bayesian methods for backpropagation networks. *Models of Neural Networks III*, pages 211–254, 1994.

D. MacKay. Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Network*, 1995.

T. Minka. *A family of approximation methods for approximate Bayesian inference*. PhD thesis, MIT, 2001.

B. Moore and C. Tan. Perceived naturalness of spectrally distored speech and music. *J. Acoust. Soc. Am.*, 114:408–419, 2003.

B. Moore and C. Tan. Development of a method for predicting the perceived naturalness of sounds subjected to spectral distortion. *J. Acoust. Soc. Am.*, 52:900–914, 2004.

M. Narendran and L. Humes. Reliability and validity of judgments of sound quality in elderly hearing aid wearers. *Ear. Hear.*, 24:4–11, 2003.

M. Nilson, S. Soli, and J. Sullivan. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.*, 95:1085–1099, 1994.

D. Preves. Expressing hearing aid noise and distortion with coherence measurements. *Asha*, 32: 56–59, 1990.

C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

S. Salzberg. On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1:317–327, 1997.

P. Souza, L. Jenstad, and K. Boike. Measuring the acoustic effects of compression and amplification on speech in noise. *J. Acoust. Soc. Am*, 119:41–44, 2006.

P. Stelmachowicz, D. Lewis, B. Hoover, and D. Keefe. Subjective effects of peak clipping and compression limiting in normal and hearing-impaired children and adults. *J. Acoust. Soc. Am*, 105:412–422, 1999.

C. Tan and B. Moore. Perception of nonlinear distortion by hearing-impaired people. *International Journal of Audiology*, 47(5):246–256, 2008.

C. Tan, B. Moore, and N. Zacharov. The effect of nonlinear distortion on the perceived quality of music and speech signals. *J. Audio. Eng. Soc.*, 51:1012–1031, 2003.

C. Tan, B. Moore, N. Zacharov, and V. Mattila. Predicting the perceived quality of nonlinearly distorted music and speech signals. *J. Audio. Eng. Soc.*, 52:699–711, 2004.