Predicting Preference Judgments of Individual Normal and Hearing-Impaired Listeners With Gaussian Processes

Perry Groot, Tom Heskes, Tjeerd M.H. Dijkstra, and James M. Kates

***Abstract*—A probabilistic kernel approach to pairwise preference learning based on Gaussian processes is applied to predict preference judgments for sound quality degradation mechanisms that might be present in a hearing aid. Subjective sound quality comparisons for 14 normal-hearing and 18 hearing-impaired subjects were used for evaluating the predictive performance. Stimuli were sentences subjected to three kinds of distortion (additive noise, peak clipping, and center clipping) with eight levels of degradation for each distortion type. The kernel approach gives a significant improvement in preference predictions of hearing-impaired subjects by individualizing the learning process. A significant difference is shown between normal-hearing and hearing-impaired subjects, because of nonlinearities in the perception of hearing-impaired subjects. In particular, hearing-impaired subjects have significant nonlinear preference judgments when making pairwise comparisons between peak clipped sentences with different clipping thresholds. The probabilistic kernel approach is shown to be robust when generalizing over distortions and over subjects.**

***Index Terms*—Bayes procedures, Gaussian process (GP), Pairwise comparisons, Subjective quality measures.**

I. 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. 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 [1], [2], vowel sounds [3], comb-filtered noise [4], psycho-acoustic and cognitive models combined with a time alignment algorithm [5], [6], audiory perception models [7], and coherence based methods [8], [9]. Tan and Moore [10] wrote several papers on the topic focusing on linear distortion, nonlinear distortion, and their combination. 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, but some systematic errors remained [9] and some model extensions have been surprisingly ineffective possibly because of random variability in the judgments of subjects [10].

Arehart *et al.* [9] collected pairwise comparisons of 14 normal-hearing and 18 hearing-impaired subjects for several sound distortions (additive noise, peak clipping, and center clipping). They 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 index (CSII) 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.

The *aim of this article* is to develop a principled modeling approach that can reproduce or predict with a maximum accuracy the observed preference judgments of an individual listener. Towards this end, we extend the analysis of Arehart *et al.* [9] by (1) fitting a model to individual listeners; (2) directly fitting the binary response data; (3) using a flexible *nonparametric* regression model based on Gaussian processes [11]; (4) devising model-independent measures for response bias and consistency. We model preferences with Gaussian processes because these can implement nonparametric nonlinear functions. This allows us to find evidence for nonlinear behaviour without assuming a specific form of nonlinearity. In [12], nonlinear and nonparametric techniques are shown to provide higher correlations with subjective speech quality and speech/noise distortions than conventional measures.

The *contribution of this article* is a significant improvement in the predictive performance for hearing-impaired individuals. Reasons for this improvement include (1) the use of individual preferences; (2) taking into account a response bias and inconsistencies in user preferences; (3) allowing for nonlinearities in the preference judgments of hearing-impaired subjects (i.e., preference judgments are modeled using nonlinear functions). As no such improvements could be made for normal-hearing subjects, this also demonstrates significant differences between the groups of normal-hearing and hearing-impaired subjects. We show that the methodology proposed is robust when generalizing over distortions and over subjects. We believe

Manuscript received January 13, 2010. The current research was funded by STW project 07605 and NWO VICI grant 639.023.604.

P.C. Groot, T. Heskes, and T.M.H. Dijkstra are with the Machine Learning Group, Intelligent Systems, Institute for Computing and Information Sciences (iCIS), Faculty of Science, Radboud University, Nijmegen, 6525 AJ Nijmegen, the Netherlands; (phone: (+31) - 24 - 3652354; e-mail: pcgroot@gmail.com). T.M.H. Dijkstra is also affiliated at Signal Processing Systems Group, Department of Electrical Engineering, Technical University of Eindhoven, 5600 MB Eindhoven, the Netherlands

J.M. Kates is with the Department of Speech Language and Hearing Sciences, University of Colorado, Boulder, Colorado. He is also affiliated at GN ReSound, Algorithm R&D department, Eindhoven, the Netherlands.

to be the first to develop a robust individualized sound quality model.

A. The structure of this article

The next three sections describes the techniques used. Section II describes Bayesian probability theory for plausible reasoning. Section III introduces the concept of Gaussian processes. Readers familiar with Bayesian probability theory and Gaussian processes may skip these sections. Section IV describes the Bayesian framework for preference learning with Gaussian processes from pairwise listening experiments. Section V describes empirical results obtained with our framework applied to data collected by Arehart *et al.* [9]. Section VI gives conclusions.

II. Bayesian Probability Theory

In this section we briefly introduce notation and implementation issues related to Bayesian inference [13]. For a tutorial application of Bayesian theory to hearing aid fitting see Dijkstra *et al.* [14]. Bayes' theorem gives rules for how probabilities can be manipulated:

$$p(\mathcal{M}|\mathcal{D},\mathcal{H}) = \frac{p(\mathcal{D}|\mathcal{M},\mathcal{H})p(\mathcal{M}|\mathcal{H})}{p(\mathcal{D}|\mathcal{H})} \quad \text{(Bayes' rule)} \qquad (1)$$

Here, $\mathcal{H}, \mathcal{D}, \mathcal{M}$ are propositions, where $\mathcal{H}$ can be interpreted as a hypothesis, $\mathcal{D}$ as the observed data, and $\mathcal{M}$ a model. The evidence is computed by integrating over all possible outcomes for the model parameters $\mathcal{M}$

$$p(\mathcal{D}|\mathcal{H}) = \int_{\mathcal{M}} p(\mathcal{D}|\mathcal{M},\mathcal{H})p(\mathcal{M}|\mathcal{H})\,\mathrm{d}\mathcal{M} \qquad (2)$$

One of the central tasks in science is to develop and compare hypotheses to account for observed data. Typically, these hypotheses include free parameters (called hyperparameters), representing, for example, some physical quantities. In a so-called type II maximum likelihood approach we choose the hyperparameters that maximize the evidence in Equation (2). In practice, there are, however, only a few cases that can be carried out analytically and one has to resort to approximation techniques for computing this integral. In our research we have used the Laplace method [15], [16], which approximates a posterior distribution over model parameters as a Gaussian. This is a reasonable choice when, as in our case, the posterior distribution is obtained by multiplying a sigmoid function with a Gaussian prior distribution [17], [18].

III. Gaussian processes

Several good introductions to the topic of Gaussian processes (GPs) are available [19], [20]. We denote vectors $\boldsymbol{x}$ and matrices $\boldsymbol{K}$ with bold-face type and their components with regular type, i.e., x_i, K_{ij}. With $\boldsymbol{x}^T$ we denote the transpose of the vector $\boldsymbol{x}$. A GP is a collection of random variables, any finite number of which have a joint Gaussian distribution. A GP is completely specified by its mean function and kernel function. In our case the random variables represent the values of a function $f(\boldsymbol{x})$ indexed by the set of sound samples $\mathcal{X}$, where we represent a sound $\boldsymbol{x}$ as a N-dimensional vector of real-valued features ($\boldsymbol{x} \in \mathbb{R}^N$).

A GP is in fact equivalent to a Bayesian interpretation of linear regression. Let

$$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}) = \sum_{n=1}^{N} w_n\phi_n(\boldsymbol{x}) \qquad (3)$$

be a linear combination of basis functions $\phi_n(\cdot)$ where $\boldsymbol{w}$ is a weight vector. If the weight vector $\boldsymbol{w}$ is drawn from a Gaussian distribution, this induces a probability distribution over functions $f(\cdot) = \boldsymbol{w}^T\boldsymbol{\phi}(\cdot)$. This distribution is a GP.

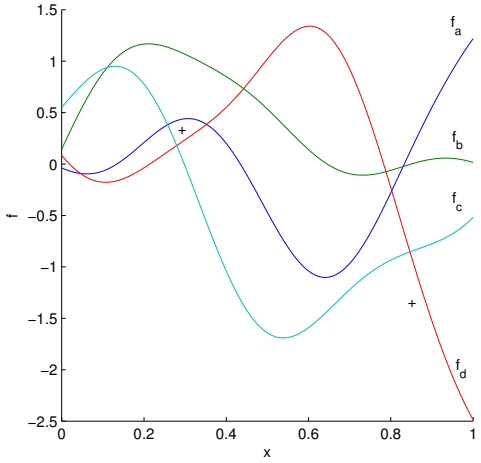


Fig. 1. Space of four functions $f_a(x)$, $f_b(x)$, $f_c(x)$, and $f_d(x)$ together with two data points (represented with $+$).

Consider, for example, the function space illustrated in Fig. 1, which consists of four different curves $f_a(x)$, $f_b(x)$, $f_c(x)$, and $f_d(x)$. If there is no reason a priori to prefer one curve over another, then each curve has an *a priori* probability of $1/4$. On observing one data point (the left-most $+$ in Fig. 1), the likelihood (or data fit) is much higher for curves f_a, f_c, and f_d than for curve f_b and their probability would be updated correspondingly. After observing a second data point (the right-most $+$ in Fig. 1), only curve f_d fits both data points well, and thus the posterior probability will have most of its mass concentrated on this curve.

We only consider GPs with a mean function equal to zero. The kernel function k (also called (co)variance function) is a function mapping two arguments into $\mathbb{R}$ that is symmetric, i.e., $k(\boldsymbol{x},\boldsymbol{x}') = k(\boldsymbol{x}',\boldsymbol{x})$, and positive definite, i.e., $\boldsymbol{z}^T\boldsymbol{K}\boldsymbol{z} > 0$ for all nonzero vectors $\boldsymbol{z}$ with real entries ($z_n \in \mathbb{R}$) and $\boldsymbol{K}$ as defined below. Intuitively, the kernel function can be viewed as a closeness or similarity measure between sounds $\boldsymbol{x}$ and $\boldsymbol{x}'$ [21], [22]. If two sounds $\boldsymbol{x}, \boldsymbol{x}'$ are similar, the function values $f(\boldsymbol{x})$ and $f(\boldsymbol{x}')$ should also be similar. The kernel function can be used to construct a covariance matrix $\boldsymbol{K}$ with respect to a set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_I\}$ by defining $K_{pq} = k(\boldsymbol{x}_p, \boldsymbol{x}_q)$. An often used kernel is the Gaussian kernel:

$$k_{\mathcal{G}}(\boldsymbol{x},\boldsymbol{x}') = \exp\left(-\frac{\ell}{2}\sum_{n=1}^{N}(x_n - x'_n)^2\right) \qquad (4)$$

where $\ell \geq 0$ is a length-scale hyperparameter, specifying how much the function can vary. A kernel function k leads to a

multivariate Gaussian prior distribution on any finite subset $\boldsymbol{f} \subseteq \{f(\boldsymbol{x}_i)\}_{\boldsymbol{x}_i \in \mathcal{X}}$ of function values

$$p(\boldsymbol{f}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\boldsymbol{K}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{f}^T \boldsymbol{K}^{-1} \boldsymbol{f}\right) \tag{5}$$

where $K_{pq} = k(\boldsymbol{x}_p, \boldsymbol{x}_q)$.

Note that we represent a function—in principle infinite dimensional—as a finite vector of function values on (a subset of) the I distinct instances in $\mathcal{X}$, i.e.,$\boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_I)]^T$. This can be done, because in the GP framework, predicting function values is computationally only dependent on the observed values, which are usually finite in number.

Furthermore, a GP is a *nonparametric* regression method. This means that one does not have to specify the form of the function $\boldsymbol{f}$ that one tries to model in advance giving more flexibility than, for example, the work of Pourmand *et al.* [23], which assumes a linear combination of radial basis functions. Empirical evidence has shown that GPs do well regardless of the input dimensionality, the degree of nonlinearity, the noise-level, and the identity of the data-family [24]. Prior assumptions about $\boldsymbol{f}$ can be specified in the kernel function k leading to different prior distributions over $\boldsymbol{f}$. The Gaussian kernel $k_{\mathcal{G}}$ allows $\boldsymbol{f}$ to be nonlinear in $\boldsymbol{x}$. The amount of smoothness, related to the length-scale ℓ of the Gaussian kernel, can be fitted automatically to the data observed (corresponding, in our case, to individual subjects).

IV. Bayesian Framework

In this section we discuss our Bayesian framework, which follows the work of Chu and Ghahramani [11], for learning subject preferences given observations from pairwise listening experiments, i.e., given two sounds, the user states his preference with respect to sound quality. In terms of Section II, our Bayesian framework consists of models $\mathcal{M} = \{f : \mathcal{X} \to \mathbb{R}\}$ and hypotheses $\mathcal{H} = \{\ell, \sigma, b\}$ explained below. The underlying idea, from economics, is that the preferences for a subject follow from an objective function (utility function) $f : \mathcal{X} \to \mathbb{R}$ evaluating sounds. Basically, this means $\boldsymbol{x}$ is preferred over $\boldsymbol{x}'$ whenever $f(\boldsymbol{x}) > f(\boldsymbol{x}')$. The hypotheses $\mathcal{H} = \{\ell, \sigma, b\}$ are a set of hyperparameters that influence the models, which we sometimes omit from formulas for the sake of readability (cf. Sections III and IV-C).

Current approaches for predicting sound quality preferences often use an absolute numeric scale (e.g., ITU-T P.835 methodology [25]) using trained individuals for rating sound quality. Our framework is motivated by the fact that in practice it is difficult to get consistent responses from *naive users* when they are asked to rate sounds *numerically* [12], [25]. Because humans *do* excel at comparing alternatives and expressing a qualitative preference for one over the other [26], [27] we use qualitative preferences to make inferences about the unobserved utility function $\boldsymbol{f}$. However, our framework can be extended to accomodate other forms of data such as, for example, the commonly used subjective mean opinion scores (MOS) [28], [29] as shown in [30].

A. Data set

Let $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_I\}$ be a set of I distinct sound samples with $\boldsymbol{x}_i \in \mathbb{R}^N$, i.e., sounds are represented as a N-dimensional vector of real-valued features. Let $\mathcal{D}$ be a set of M observed pairwise preference comparisons over instances in $\mathcal{X}$, i.e.,

$$\mathcal{D} = \{\langle \boldsymbol{x}_{i_1(m)}, \boldsymbol{x}_{i_2(m)}, d_m \rangle \,|\, 1 \le m \le M, d_m \in \{-1, 1\}\}$$

with $i_1, i_2 : \{1, \ldots, M\} \to \{1, \ldots, I\}$ index functions such that $\boldsymbol{x}_{i_1(m)}, \boldsymbol{x}_{i_2(m)}$ denote, respectively, the first and second sound sample presented in the mth listening experiment, and d_m represents the subject's preference with respect to listening experiment m. We use $d_m = 1$ when $\boldsymbol{x}_{i_1(m)}$ is preferred over $\boldsymbol{x}_{i_2(m)}$ and $d_m = -1$ otherwise.

B. Gaussian process prior $p(\boldsymbol{f})$

As explained in Section III, we define a prior over functions by assuming that function values $\{f(\boldsymbol{x}_i)\}$ are a realization of random variables in a zero-mean Gaussian process, specified by a covariance matrix or kernel function. Here we use the Gaussian kernel (4) and the linear kernel [11], defined as

$$k_{\mathcal{L}}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{n=1}^{N} x_n x'_n \tag{6}$$

The Gaussian kernel is a common choice, whereas the linear kernel corresponds to Bayesian linear regression.

C. Likelihood function $p(\mathcal{D}|\boldsymbol{f})$

The likelihood function is a mapping denoting how likely the preference data $\mathcal{D}$ are given function $\boldsymbol{f}$. We make the standard independence assumption that the likelihood can be evaluated on individual observations

$$p(\mathcal{D}|\boldsymbol{f}) = \prod_{m=1}^{M} p(\langle \boldsymbol{x}_{i_1(m)}, \boldsymbol{x}_{i_2(m)}, d_m \rangle | f(\boldsymbol{x}_{i_1(m)}), f(\boldsymbol{x}_{i_2(m)}))$$

Under ideal circumstances, i.e., when subjects can make perfect statements about their preferences, the likelihood function could be defined as

$$p_{ideal}(\langle \boldsymbol{x}_{i_1(m)}, \boldsymbol{x}_{i_2(m)}, d_m \rangle | f(\boldsymbol{x}_{i_1(m)}), f(\boldsymbol{x}_{i_2(m)})) = \begin{cases} 1 & \text{if } f(\boldsymbol{x}_{i_1(m)}) > f(\boldsymbol{x}_{i_2(m)}) \text{ and } d_m = 1 \\ 1 & \text{if } f(\boldsymbol{x}_{i_1(m)}) \le f(\boldsymbol{x}_{i_2(m)}) \text{ and } d_m = -1 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

In practice, however, subjects are sometimes inconsistent and could be influenced by the order in which we present sound samples. In order to deal with this less than ideal situation we augment the likelihood function with additional variables. Analogously to Chu and Ghahramani [11], we assume that for each sound sample presented, its evaluation is contaminated with Gaussian noise $\delta_1, \delta_2 \sim \mathcal{N}(0, \sigma^2)$. Furthermore, because subjects are forced to choose a preference over sounds even though they may not hear any difference in sound quality, we assume that subjects have a response bias b that depends on the order of sound samples presented. We noticed that adding a response bias in the context of listening

experiments, significantly improves upon the likelihood function proposed by Chu and Ghahramani [11]. The augmented likelihood function then becomes

$$\begin{aligned} &p(\langle \boldsymbol{x}_{i_1(m)}, \boldsymbol{x}_{i_2(m)}, 1\rangle | f(\boldsymbol{x}_{i_1(m)}), f(\boldsymbol{x}_{i_2(m)})) \\ &\quad = p(f(\boldsymbol{x}_{i_1(m)}) + \delta_1 > f(\boldsymbol{x}_{i_2(m)}) + b + \delta_2) \end{aligned} \tag{8}$$

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.

Equation (8) can be rewritten in terms of a cumulative Gaussian [31]

$$p(\langle \boldsymbol{x}_{i_1(m)}, \boldsymbol{x}_{i_2(m)}, d_m\rangle | f(\boldsymbol{x}_{i_1(m)}), f(\boldsymbol{x}_{i_2(m)})) = \Phi(z_m) \tag{9}$$

with $\Phi(z) = \int_{-\infty}^{z} \mathcal{N}(x; 0, 1)\,\mathrm{d}x$, $d_m \in \{-1, 1\}$, and

$$z_m = \frac{d_m(f(\boldsymbol{x}_{i_1(m)}) - f(\boldsymbol{x}_{i_2(m)}) - b)}{\sqrt{2}\sigma}, \tag{10}$$

which is a common choice of response function in discrete choice methods and is also known as the probit function [32]. Other response functions, such as the Bradley-Terry-Luce (BTL) model [33], [34], can also be incorporated as a likelihood function in the current framework. As the bias and variability in judgment of a subject are explicitly written as a parameter of the likelihood function they can be optimized with respect to the data.

D. Posterior probability $p(\boldsymbol{f}|\mathcal{D})$ and model selection

The posterior probability can now be computed by applying Bayes' rule (1)

$$\begin{aligned} p(\boldsymbol{f}|\mathcal{D}, \mathcal{H}) &= \frac{p(\boldsymbol{f}|\mathcal{H})p(\mathcal{D}|\boldsymbol{f}, \mathcal{H})}{p(\mathcal{D}|\mathcal{H})} \\ &= \frac{p(\boldsymbol{f}|\mathcal{H})}{p(\mathcal{D}|\mathcal{H})} \prod_{m=1}^{M} \Phi(z_m) \end{aligned} \tag{11}$$

with $\mathcal{H} = \{\ell, \sigma, b\}$ the hyperparameters of the model.

As a Gaussian prior multiplied by a cumulative Gaussian (the likelihood function) does not lead to a closed form formula, approximations are needed to compute the posterior probability. We use the Laplace approximation [16]. Model selection is done by maximum likelihood estimation of the hyperparameters $\mathcal{H} = \{\ell, \sigma, b\}$, which we compute using gradient descent. For details see our technical report [31].

V. Empirical Results

A. Methodology

1) Data set: We use data from Arehart *et al.* [9], 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 as cochlear impairment was consistent with the results of an audiometric evaluation during their initial visit. The stimulus presented were two sets (one male, one female talker) of concatenated sentences from the hearing-in-noise-test (HINT) [35]. 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 stimulus was subjected to $I = 24$ 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—*a full orthogonal design*. In total $J = (2+3+3)\cdot 72 = 576$ paired comparisons were collected for each subject.

In our approach to model and predict the preference judgment of listeners with respect to 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.* [9] 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 gives a procedure to take into account audibility factors [36]. Each degraded test sentence was divided into three amplitude regions, at or above the root-mean squared sentence level (RMS), 0-10 dB below the RMS level, and 10-30 dB below the RMS level. The signal envelope was computed using a Hamming-windowed segment size of 16 ms. The procedure results for each sound sample in three features, namely CSII_{Low}, CSII_{Mid}, and $\text{CSII}_{\text{High}}$, for the low-, mid-, and high-level portions of the speech. We use the CSII approach to represent the pairwise preference data of Arehart *et al.* [9], i.e., the data consists of pairwise experiments $\langle \boldsymbol{x}, \boldsymbol{x}', d\rangle$ of two sound samples $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ and subject decision $d \in \{1, -1\}$ denoting whether $\boldsymbol{x}$ is preferred over $\boldsymbol{x}'$ or $\boldsymbol{x}'$ is preferred over $\boldsymbol{x}$, respectively. Each sound sample is represented by three features CSII_{Low}, CSII_{Mid}, and $\text{CSII}_{\text{High}}$, which can take values in $[0, 1]^3$.

2) K-fold cross-validation: When learning a model from data, the goal is to make good predictions for data *not yet observed*. Therefore, one has to take care that the model generalizes well to situations not yet presented and does not *overfit* the gathered data. Here we use cross-validation, which splits the data into disjoint subsets used for training and testing of the model, to validate that there is no overfitting to the training data. Note that we do *not* use cross-validation to search for optimal hyperparameters as this is done by maximizing the evidence (cf. Section IV-D).

In K-fold cross-validation [37], 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 folds (10-fold cross-validation is a common choice and is also used here). For each fold $k \in \{1, \ldots, K\}$ the classifier is trained on the data set $\mathcal{D} \setminus \mathcal{D}_k$ and tested on the data set $\mathcal{D}_k$. Each sample is thus used for training and used exactly once for testing.

The K results from the folds are combined to produce a single cross-validation estimate of accuracy. This is the number of correct classifications, divided by the number of instances $|\mathcal{D}| = J = 576$. Formally, let $\mathcal{C}(\mathcal{D}, \langle \boldsymbol{x}, \boldsymbol{x}'\rangle)$ be the classifier that returns a label for samples $\langle \boldsymbol{x}, \boldsymbol{x}'\rangle$ when trained on $\mathcal{D}$. Let

$\mathcal{D}_{(j)}$ be the test set that contains instance $\langle \boldsymbol{x}_{i_1(j)}, \boldsymbol{x}_{i_2(j)}, d_j \rangle$, then the *cross-validation estimate of accuracy* is defined as

$$acc_{CV} = \frac{1}{J} \sum_{j=1}^{J} \delta(\mathcal{C}(\mathcal{D} \setminus \mathcal{D}_{(j)}, \langle \boldsymbol{x}_{i_1(j)}, \boldsymbol{x}_{i_2(j)} \rangle), d_j) \quad (12)$$

where $\delta(p, q) = 1$ iff $p = q$ and $\delta(p, q) = 0$ iff $p \neq q$. The prediction error (PE) is $1 - acc_{CV}$.

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 [38]–[40]. 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. McNemar's test disregards the easy to classify cases and focuses on the hard to classify cases. McNemar's test was also found to be one of the best statistical tests for comparing supervised classification algorithms in an empirical evaluation by Dietterich [39].

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 [39]

J_{00}: Number of examples misclassified by both $\hat{f}_A$ and $\hat{f}_B$	J_{01}: Number of examples misclassified by $\hat{f}_A$ but not by $\hat{f}_B$
J_{10}: Number of examples misclassified by $\hat{f}_B$ but not by $\hat{f}_A$	J_{11}: Number of examples misclassified by neither $\hat{f}_A$ nor $\hat{f}_B$.

where $J = J_{00}+J_{01}+J_{10}+J_{11}$ is the total number of samples (pairwise comparisons).

Under the null hypothesis that both algorithms perform equally well, the two algorithms should have the same error rate, which means that $J_{01} = J_{10}$. 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

J_{00}	$(J_{01} + J_{10})/2$
$(J_{01} + J_{10})/2$	J_{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:

$$\frac{(|J_{01} - J_{10}| - 1)^2}{J_{01} + J_{10}} \quad (13)$$

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.

B. Classifier comparison results

In this section we compare three classifiers for the pairwise comparison data of Arehart *et al.* [9]. The first classifier is the Q_3 metric reported by Arehart *et al.* [9], which is a minimum mean-squared error fit to the normal-hearing subjects' quality ratings and is given by

$$Q_3 = \frac{1}{1 + e^{-c}} \quad (14)$$

with

$$c = -4.56 + 2.41 \cdot \text{CSII}_{\text{Low}} + 2.16 \cdot \text{CSII}_{\text{Mid}} + 1.73 \cdot \text{CSII}_{\text{High}}$$

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 CSII 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 IV we trained two other classifiers. A Gaussian process with a linear kernel (LK) and a Gaussian process with a Gaussian kernel (GK). We use GK_0 to denote a GK model where the bias hyperparameter $b = 0$ is fixed, i.e., no bias. The linear kernel corresponds to probit regression for each *individual* subject. The Gaussian process with linear kernel would give similar results to the Q_3 metric if we would apply it on the pooled data set consisting of all the normal-hearing subject data. The order of the pairwise comparisons in each data file was randomized before the data file was split into 10 equal sized blocks in order to train both kernels using 10-fold cross-validation, i.e., both kernels used the same splits of the data files. The classifier results were compared to each other using the McNemar test (cf. Section V-A2) and are shown in Table I.

The first column is an identifier indicating the subject. The top half (from 'nh1' to 'nh14') are the 14 normal-hearing subjects. The bottom half (from 'hi1' to 'hi18') are the 18 hearing-impaired subjects. The second and third column report on the response bias and the level of inconsistency of the listener (discussed in the next paragraph). Columns 4, 5, and 6 report the prediction error (PE) as well as the correlation coefficient r for the three classifiers on each data set corresponding to a single subject. For the correlation coefficient we follow [9] by computing individual preference scores (the proportion of the times a signal degradation condition was preferred to all the other conditions) based on the data and the model predictions. The last four columns report the comparison results between the classifiers using McNemar's test. Here, 'p' is the reported p-value, 'J_{10}' is the number of examples correctly predicted by the first named classifier, but wrongly predicted by the second named classifier, and 'J_{01}' is the number of examples 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 'overall', i.e., average percentages and prediction errors, and overall McNemar test results (i.e., the p-value is based on the sum of J_{10} and J_{01} scores).

The second column reports the percentage of biased pairs of examples, i.e., with $\boldsymbol{x} \neq \boldsymbol{x}'$ the percentage of pairs such

TABLE I

CLASSIFIER COMPARISON RESULTS FOR NORMAL-HEARING SUBJECTS (FROM 'NH1' TO 'NH14') AND HEARING-IMPAIRED SUBJECTS (FROM 'HI1' TO 'HI18'). COLUMNS 2 AND 3 REPORT DATA STATISTICS. COLUMNS 4, 5, AND 6 REPORT THE PREDICTION ERROR (PE) AND CORRELATION COEFFICIENT r OF THE THREE CLASSIFIERS COLUMNS 7, 8, 9, AND 10 REPORT COMPARISONS BETWEEN CLASSIFIERS USING MCNEMAR'S TEST WITH 'P' THE P-VALUE, 'J_{10}' THE NUMBER OF EXAMPLES CORRECTLY CLASSIFIED BY THE FIRST BUT NOT BY THE SECOND NAMED CLASSIFIER, AND 'J_{01}' THE NUMBER OF EXAMPLES CORRECTLY CLASSIFIED BY THE SECOND BUT NOT BY THE FIRST NAMED CLASSIFIER.

Subj.	%	%	Q_3		LK		GK		Q_3 vs LK			Q_3 vs GK			LK vs GK			GK_0 vs GK		
name	bias	incons.	PE	r	PE	r	PE	r	p	J_{10}	J_{01}	p	J_{10}	J_{01}	p	J_{10}	J_{01}	p	J_{10}	J_{01}
nh1	21.7	0.11	0.15	0.94	0.14	0.94	0.15	0.95	0.81	8	10	0.68	13	10	0.23	8	3	1.00	8	7
nh2	14.5	0.21	0.13	0.97	0.11	0.97	0.11	0.97	0.14	14	24	0.07	15	28	0.58	5	8	0.44	18	24
nh3	13.4	0.27	0.12	0.95	0.12	0.94	0.12	0.96	0.86	16	14	0.88	22	20	0.86	16	16	0.18	18	10
nh4	10.1	0.14	0.12	0.94	0.12	0.94	0.12	0.95	1.00	20	19	0.55	25	20	0.39	8	4	0.31	28	20
nh5	15.9	0.21	0.14	0.93	0.15	0.93	0.14	0.94	0.15	16	8	0.86	15	15	0.17	9	17	0.65	11	8
nh6	17.4	0.20	0.14	0.92	0.14	0.93	0.13	0.94	0.86	14	16	0.57	22	27	0.74	16	19	0.45	19	25
nh7	18.5	0.61	0.18	0.86	0.18	0.86	0.16	0.92	1.00	27	28	0.42	34	42	0.42	24	31	0.44	34	27
nh8	22.8	1.28	0.21	0.89	0.21	0.89	0.19	0.92	0.88	22	22	0.19	24	35	0.14	17	28	1.00	22	23
nh9	14.1	0.32	0.16	0.90	0.16	0.89	0.17	0.91	0.87	19	21	1.00	19	18	0.78	26	23	0.23	8	3
nh10	15.9	0.33	0.14	0.95	0.14	0.95	0.13	0.96	1.00	11	10	0.58	13	17	0.40	9	14	0.45	6	10
nh11	18.1	0.68	0.16	0.94	0.18	0.93	0.17	0.95	0.05	22	10	0.17	27	17	0.84	11	13	0.06	17	7
nh12	19.9	0.74	0.15	0.91	0.15	0.92	0.13	0.96	1.00	19	18	0.20	25	36	0.10	17	29	0.29	18	26
nh13	22.5	0.27	0.15	0.95	0.15	0.95	0.16	0.94	0.87	19	19	1.00	20	19	1.00	5	4	0.85	12	14
nh14	18.5	0.33	0.14	0.98	0.12	0.98	0.12	0.98	0.16	12	21	0.26	15	23	1.00	6	5	0.23	9	16
overall	17.4	0.41	0.15	0.93	0.15	0.93	0.14	0.95	1.00	239	240	0.14	289	327	0.07	177	214	0.78	227	220
hi1	37.7	0.82	0.26	0.81	0.19	0.84	0.11	0.98	0.00	56	94	0.00	29	115	0.00	17	65	0.00	24	54
hi2	21.7	0.57	0.16	0.95	0.11	0.98	0.11	0.99	0.01	22	46	0.00	20	45	1.00	13	14	0.05	13	26
hi3	18.8	0.82	0.18	0.87	0.16	0.89	0.16	0.93	0.27	28	38	0.28	37	48	1.00	24	25	0.69	30	26
hi4	22.8	0.43	0.18	0.94	0.15	0.95	0.14	0.96	0.03	30	51	0.00	24	52	0.32	15	22	0.01	20	41
hi5	32.2	0.84	0.20	0.96	0.14	0.96	0.14	0.97	0.00	33	64	0.00	33	64	0.80	8	8	0.01	33	58
hi6	27.9	1.43	0.24	0.78	0.23	0.82	0.21	0.90	0.34	31	40	0.06	37	56	0.22	22	32	0.49	41	34
hi7	18.1	0.44	0.14	0.96	0.11	0.97	0.11	0.98	0.04	20	36	0.02	17	35	0.79	6	8	0.07	10	21
hi8	21.7	1.23	0.18	0.92	0.17	0.93	0.15	0.98	0.88	22	24	0.11	22	35	0.14	17	28	0.11	18	30
hi9	18.1	0.59	0.15	0.92	0.12	0.96	0.12	0.98	0.03	15	31	0.01	15	35	0.62	16	20	0.00	3	17
hi10	15.6	0.26	0.14	0.96	0.12	0.97	0.11	0.99	0.01	7	21	0.01	12	31	0.42	10	15	0.54	10	14
hi11	29.3	0.22	0.16	0.95	0.12	0.96	0.10	0.98	0.01	22	44	0.00	18	53	0.03	9	22	0.00	16	48
hi12	35.5	0.31	0.19	0.88	0.14	0.91	0.09	0.99	0.00	28	56	0.00	18	79	0.00	15	48	0.00	17	56
hi13	35.9	2.00	0.24	0.94	0.22	0.94	0.18	0.97	0.27	44	56	0.00	45	78	0.00	15	36	0.37	36	45
hi14	25.0	0.66	0.18	0.90	0.13	0.96	0.11	0.97	0.00	21	52	0.00	16	58	0.03	6	17	0.07	15	28
hi15	14.1	0.22	0.13	0.98	0.14	0.98	0.12	0.98	0.80	32	29	0.72	33	37	0.19	7	14	0.64	23	19
hi16	30.1	2.05	0.31	0.64	0.23	0.89	0.20	0.96	0.00	63	109	0.00	45	108	0.06	27	44	0.45	6	10
hi17	31.9	3.04	0.24	0.88	0.22	0.96	0.22	0.96	0.28	37	48	0.19	36	49	0.75	4	6	0.77	22	25
hi18	30.4	0.49	0.18	0.97	0.15	0.97	0.15	0.97	0.05	19	34	0.04	24	42	0.69	11	14	0.19	23	34
overall	25.9	0.91	0.19	0.90	0.16	0.94	0.14	0.97	0.00	530	873	0.00	481	1020	0.00	242	438	0.00	360	586

that $\langle \boldsymbol{x}, \boldsymbol{x}', d\rangle$ and $\langle \boldsymbol{x}', \boldsymbol{x}, d\rangle$ holds for $d \in \{1, -1\}$, giving a base rate of 50%. The third column reports a consistency check *independent from the response bias*. For this, we counted for all quadruples $A, B, C, D \in \mathcal{X}$ distinct whether the subject's preferences $d_m \in \{-1, 1\}$ in $\langle A, B, d_1\rangle$, $\langle C, B, d_2\rangle$, $\langle A, D, d_3\rangle$, $\langle C, D, d_4\rangle$ are consistent with some total ordering over $\langle A, B, C, D\rangle$. Note, that only $\langle d_1, d_2, d_3, d_4\rangle \in \{\langle -1, 1, 1, -1\rangle, \langle 1, -1, -1, 1\rangle\}$ does not lead to a total order over $\langle A, B, C, D\rangle$ giving a base rate of 12.5%.

The second column shows that normal-hearing subjects have a lower response bias than hearing-impaired subjects (mean 17.4% versus 25.9%) and there is less variability within the group (standard deviation 0.95% versus 1.69%). 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). The low percentage of inconsistency comes from the large number of obvious preference relations and a low base rate of 12.5% for a random guesser.

The tenth column shows the effect of incorporating a bias parameter in the likelihood function using a GK classifier (i.e., GK_0 stands for a GK model with fixed $b = 0$ meaning no bias). Incorporating a response bias significantly improved the GK model for hearing-impaired subjects, i.e., for each of the 18 hearing-impaired subjects the McNemar test showed six significant differences ($p < 0.05$) in favour of a GK model with a response bias when compared with a GK model without a response bias.

Looking at the results for the normal-hearing subjects, we see that the classification performances of the three classifiers Q_3, LK, and GK are very similar. Sometimes one classifier performs better than another, sometimes worse. The classification performance of one classifier, however, is never significantly better than another classifier as all reported p-values are greater than 0.05. This changes, however, when we consider the classification performance on the data corresponding to the hearing-impaired subjects. The Q_3 metric is significantly outperformed by the LK and GK classifiers on 11 and 13 subjects respectively. The LK classifier is again significantly outperformed by the GK classifier on 5 subjects. 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. Furthermore, for normal-hearing data the LK classifier is basically the same as the Q_3 classifier except

applied to individual subjects instead of pooled data.

It follows from these results that the prediction of preference judgments for hearing-impaired subjects with respect to speech 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; This follows from the 'Q_3 vs LK' column which compares a linear model on pooled data with a linear model on individual data), and by (2) allowing nonlinear relationships in the decision model (the latter follows from the 'LK vs GK' column which compares a linear model with a nonlinear 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 shows 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.

C. Validation of coherence speech intelligibility index

One of the questions that may arise from the results reported in Table I and Fig. 2 is that the nonlinear behavior in perception of hearing-impaired subjects is a result of the use of the CSII measure that transforms the sound features based on the audiogram of the subject. In order to validate the CSII 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 CSII which incorporates 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 II, which shows that the CSII, which incorporates the audiogram, significantly improved the overall results as well as the individual results of three hearing-impaired subjects. This validates the approach followed of using the CSII measure which incorporates the audiogram of the subject. Note that this approach differs from related work of Tan and Moore [10] who do not subject their distorted signals to the frequency-dependent amplification prescribed by the Cambridge-formula.

D. Nonlinearity

To demonstrate the nonlinear perception in hearing-impaired subjects we show in Fig. 2 middle panel the elicited utility function of a typical normal-hearing subject (top row) and a hearing-impaired subject (bottom row) for which we obtained a significant improvement in predicting preference judgments because of a nonlinear utility function. As the CSII_{Mid} features were highly correlated with the CSII_{Low} and $\text{CSII}_{\text{High}}$ features we used linear regression for each subject to effectively reduce our graph to 3-dimensions. For example, for subject 'hi12' we obtained the following linear regression relation ($R^2 = 0.68$):

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

Fig. 2 shows the hyperplane in terms of CSII_{Low}, $\text{CSII}_{\text{High}}$, and $\text{CSII}_{\text{Mid}}{}'$ features, and the samples projected on the contour plot of the utility function.

TABLE II
PREDICTION ERROR (PE) OF A GAUSSIAN PROCESS WITH GAUSSIAN KERNEL FOR HEARING-IMPAIRED SUBJECTS USING COHERENCE BASED FEATURES (GK) AND FEATURES USED BY NORMAL-HEARING SUBJECTS (GK-NH). COMPARISONS ARE MADE USING MCNEMAR'S TEST WITH 'P' THE P-VALUE, 'J_{10}' THE NUMBER OF EXAMPLES CORRECTLY CLASSIFIED BY THE FIRST BUT NOT BY THE SECOND NAMED CLASSIFIER, AND 'J_{01}' THE NUMBER OF EXAMPLES CORRECTLY CLASSIFIED BY THE SECOND BUT NOT BY THE FIRST NAMED CLASSIFIER.

Subj. name	GK PE	GK-nh PE	GK vs GK-nh p	J_{10}	J_{01}
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
overall	0.14	0.15	0.00	358	251

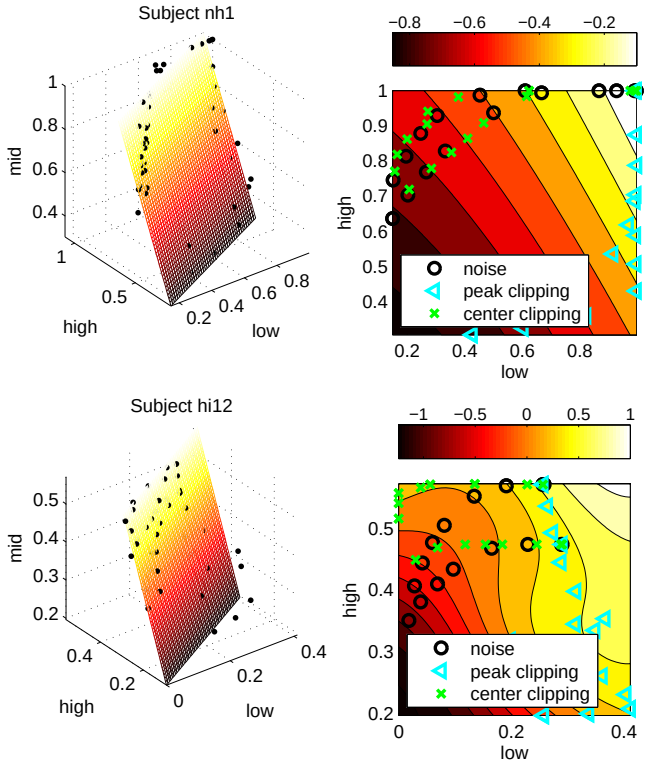


Fig. 2. Utility function elicited for a normal-hearing subject 'nh1' (top) and hearing-impaired subject 'hi12' (bottom). Left: hyperplane formed by linear regression of CSII_{Mid} in terms of CSII_{Low} and $\text{CSII}_{\text{High}}$ features. Right: contour plot with sound samples distorted with noise (circle), peak clipping (triangle), and center clipping (cross). The noise levels of the sound samples are lowest in the upper right corner and increase when moving into the direction of the lower left corner.

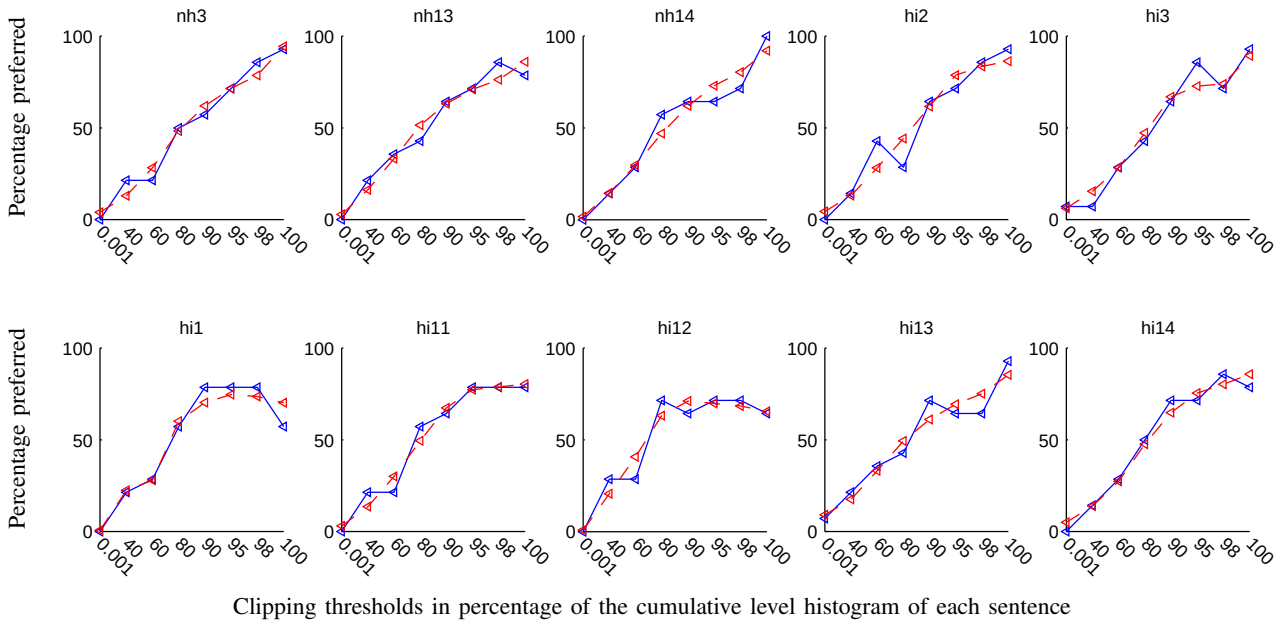


Fig. 3. Comparison *within* the peak clipping degradation for ten subjects. The solid line is calculated from the data whereas the dashed line is the model prediction using a Gaussian process with a nonlinear Gaussian kernel (by summing up predictive probabilities). The y-axis represents the fraction (percentage) of times that a particular distortion level is preferred over all other distortion levels. The x-axis denotes the clipping threshold used in terms of the percentage of the cumulative level histogram of each sentence. The effect of peak clipping is reduced as the clipping threshold is increased. With respect to improved performance in preference prediction when comparing a linear and nonlinear model, the top row shows the least significant results and the bottom row shows the most significant results (cf. Table I).

To get more insight into the observed nonlinear behaviour we went back to the original data set and counted for each of the three distortion types per level how often it was preferred when compared to the same distortion type, i.e., comparisons are *within* a distortion type. The results for peak clipping are shown in Fig. 3. The top row corresponds to the five least significant differences and the bottom row to the five most significant differences in Table I, which reports the comparison between a linear and nonlinear model. Results for center clipping and noise are not shown, because preferences for all normal-hearing and hearing-impaired listeners were quite linear. From Fig. 3 it can be observed that the preferences for peak clipping are not always linear with respect to the distortion level. The preferences for peak clipping show a drop at the end of our range. The Q_3 metric of Arehart *et al.* [9] would predict a line from the left bottom corner to the right upper corner for each individual, whereas the GP clearly fits a nonlinear smooth curve to the data of each individual.

Note that in Fig. 3 the least significant results (top row) contain both normal-hearing and hearing-impaired subjects, whereas the most significant results (bottom row) contain only hearing-impaired subjects. From these results, it seems that some hearing-impaired subjects have difficulty in hearing differences between peak clipped speech at the end of our range, which leads to nonlinear behaviour in their preference decisions. We also noted some drop in the middle of our range for center clipped sentences, but did not note any clear differences between the least and most significant results.

E. Generalization over distortions and over subjects

So far, we have shown a significant improvement in predictive performance of sound quality perception of hearing-impaired subjects. Here, we take a look at the generalization behaviour of the GP model over distortions and over subjects.

In the case of generalizing over distortions we split the data per subject in a training set (256 comparisons) and test set (320 comparisons). The training set consisted of all pairwise comparisons of two distortions (e.g., noise and peak-clipping) and the test set of all pairwise comparisons containing the left out distortion (e.g., center-clipping). Per subject, this resulted in three experiments for each possible division of distortions in training and test set. The results are shown in Table III. Each column heading shows the distortion put in the test set and the prediction error on this set per subject. Per subject, the coefficients in the Q_3 metric were re-fitted using the 256 comparisons in the training data. For the normal-hearing subjects both the Q_3 metric and the GP model perform equally well, except that the GP model tends to overfit to the training data when the peak clipping distortion is left out of the training set. This is exactly what one should expect when taking Fig. 2 into account as the features of the signals distorted with noise are similar to the features of the signals distorted with center-clipping, but quite different from the features of the signals distorted with peak-clipping. The same remarks hold for the hearing-impaired subjects, except that the GP model performs better than Q_3 when the comparisons involving noise are left out of the training data.

In the case of generalizing over subjects, per subject we pooled all data of all other subjects into one big training set

TABLE III
CLASSIFICATION RESULTS FOR THE GENERALIZATION OVER DISTORTIONS. THE PREDICTION ERROR IS REPORTED IN COLUMNS 2, 3, AND 4 FOR THE (RE-FITTED) Q_3 CLASSIFIER, IN 5, 6, AND 7 FOR THE GAUSSIAN PROCESS WITH GAUSSIAN KERNEL FUNCTION. THE COLUMN HEADING DENOTES THE DISTORTION PRESENT IN THE TEST SET.

Subj. name	Q_3			GK		
	Noise	PClip	CClip	Noise	PClip	CClip
nh1	0.17	0.18	0.16	0.20	0.46	0.19
nh2	0.16	0.18	0.15	0.13	0.22	0.12
nh3	0.15	0.17	0.13	0.16	0.22	0.15
nh4	0.15	0.16	0.17	0.16	0.14	0.17
nh5	0.18	0.16	0.18	0.17	0.24	0.21
nh6	0.18	0.18	0.18	0.14	0.38	0.20
nh7	0.24	0.22	0.20	0.24	0.37	0.20
nh8	0.26	0.22	0.23	0.25	0.24	0.22
nh9	0.23	0.19	0.18	0.22	0.23	0.18
nh10	0.18	0.17	0.17	0.19	0.26	0.17
nh11	0.21	0.19	0.18	0.22	0.35	0.17
nh12	0.18	0.18	0.18	0.17	0.33	0.16
nh13	0.18	0.16	0.18	0.19	0.22	0.17
nh14	0.15	0.19	0.14	0.14	0.14	0.12
mean	0.19	0.18	0.17	0.18	0.27	0.17
hi1	0.28	0.44	0.22	0.28	0.26	0.28
hi2	0.16	0.31	0.18	0.11	0.19	0.14
hi3	0.25	0.26	0.18	0.22	0.51	0.24
hi4	0.21	0.25	0.19	0.17	0.19	0.18
hi5	0.20	0.27	0.21	0.14	0.14	0.15
hi6	0.32	0.31	0.23	0.27	0.55	0.27
hi7	0.16	0.24	0.13	0.12	0.38	0.15
hi8	0.21	0.23	0.20	0.20	0.30	0.20
hi9	0.18	0.29	0.15	0.15	0.46	0.19
hi10	0.15	0.18	0.15	0.14	0.32	0.12
hi11	0.20	0.22	0.15	0.12	0.20	0.12
hi12	0.14	0.28	0.19	0.16	0.41	0.20
hi13	0.26	0.28	0.25	0.19	0.31	0.21
hi14	0.21	0.31	0.14	0.16	0.30	0.15
hi15	0.13	0.25	0.14	0.13	0.13	0.11
hi16	0.30	0.54	0.32	0.22	0.54	0.38
hi17	0.24	0.38	0.28	0.23	0.23	0.22
hi18	0.17	0.24	0.18	0.14	0.26	0.14
mean	0.21	0.29	0.19	0.18	0.32	0.19

of 17 856 pairwise comparisons while the data of the subject itself was taken to be the test set consisting of 576 pairwise comparisons. The results are shown in Table IV. Comparing the results in Table IV with the sixth column (GK PE) in Table I, we observe that using pooled data from other subjects is always better than using individual subject data in the case of normal-hearing subjects, but the results are the other way around in the case of hearing-impaired subjects. This can be explained by the similarity between normal-hearing subjects and the diversity between hearing-impaired subjects. The modeling could be further improved by considering multi-task learning [41], [42]. Multi-task learning does not pool any data but uses the data of individuals to regularize each other. It is possible to learn the correlation among normal and hearing-impaired listeners jointly, rather than build models separately for each subject.

F. Generalization to different feature representations

The results presented so far assume that sounds are represented as a 3-dimensional real-valued vector. One might expect that the predictive accuracy could be further improved by adding more relevant features. In order to investigate the influence of the feature representation on the outcome of our methods we modified the MATLAB code used to generate the three-level coherence based speech intelligibility index (CSII) measure. We made the following changes: (1) we included a delay compensation filter in the coherence calculation; (2) we included 8 kHz hearing loss (from Table 1 in Arehart *et al.* [9]); (3) we picked level boundaries based on percentiles, not absolute values. For example, the 0th, 50th, and 100th percentiles were used to generate a 2-dimensional feature representation. We performed similar predictive experiments as reported in Table I using the Gaussian kernel (GK) and 1-5 features. Using 1 feature resulted in significantly worse predictive outcomes whereas using 2-5 features had *no significant effect*. Whether the results are robust using other feature representations is, however, beyond the scope of this paper.

TABLE IV
CLASSIFICATION RESULTS FOR THE GENERALIZATION OVER SUBJECTS. THE PREDICTION ERROR IS REPORTED IN COLUMNS 2 AND 3 FOR THE (RE-FITTED) Q_3 CLASSIFIER AND THE GAUSSIAN PROCESS WITH GAUSSIAN KERNEL FUNCTION, RESPECTIVELY. EACH ROW IN THE FIRST COLUMN DENOTES THE SUBJECT GENERALIZED OVER (I.E., TEST SET).

Subj. name	Q_3 Listener	GK Listener
nh1	0.15	0.13
nh2	0.13	0.09
nh3	0.13	0.09
nh4	0.13	0.07
nh5	0.15	0.10
nh6	0.16	0.12
nh7	0.19	0.14
nh8	0.20	0.17
nh9	0.16	0.14
nh10	0.15	0.11
nh11	0.16	0.15
nh12	0.15	0.11
nh13	0.16	0.14
nh14	0.13	0.11
mean	0.15	0.12

Subj. name	Q_3 Listener	GK Listener
hi1	0.26	0.23
hi2	0.16	0.21
hi3	0.18	0.14
hi4	0.18	0.15
hi5	0.19	0.17
hi6	0.25	0.22
hi7	0.14	0.15
hi8	0.18	0.15
hi9	0.15	0.15
hi10	0.13	0.19
hi11	0.18	0.13
hi12	0.19	0.44
hi13	0.23	0.30
hi14	0.18	0.14
hi15	0.13	0.18
hi16	0.34	0.32
hi17	0.25	0.26
hi18	0.16	0.14
mean	0.19	0.20

VI. CONCLUSIONS

This study began with the premise that the preference judgments of subjects with respect to sounds is a central issue in the development of hearing aids and other communication devices. Methods for accurately predicting subject's preference judgments would advance their development.

In this study we advocated a Bayesian framework using GPs, which can take into account response biases and inconsistencies in user preferences and is robust when generalizing over distortions and over listeners. We demonstrated that predicting the preference judgments of a hearing-impaired subject can significantly be improved by (1) learning from the subject's individual preferences, and (2) allowing nonlinearities in the decision model. No such improvements could be made for normal-hearing subjects, indicating significant differences between both groups of subjects. One should thus be careful generalizing models learned from/fitted on normal-hearing subjects to hearing-impaired subjects.

The proposed methodology has focused mainly on the prediction of preference judgments with respect to sound quality, but can be generalized to include other observations such as domestic information like age and gender, or generalized to other quality measures such as mean opinion scores and intelligibility. The statistical model for intelligibility, however, would be more complex than the model for quality. As intelligibility is typically measured in a 'repeat-the-sentence paradigm' as in the HINT test, the intelligibility score is the number of correctly understood sentences (say s out of S sentences). This can be modeled with a generalization of the Gaussian process probit model that we proposed, but the mathematics would be more involved.

ACKNOWLEDGMENT

We thank Adriana Birlutiu, Dr. Bert de Vries, and Iman Mossavat for earlier discussions. Furthermore, we thank the anonymous reviewers for their comments on an earlier draft.

REFERENCES

[1] 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.*, vol. 49, pp. 1011–1048, 2001.

[2] ——, "Multitone testing of sound system components: Some results and conclusions, Part 2: Modeling and application," *J. Audio. Eng. Soc.*, vol. 49, pp. 1181–1192, 2001.

[3] H. Levitt, E. Cudahy, H. Hwang, E. Kennedy, and C. Link, "Towards a general measure of distortion," *J. Rehab. Res. Dev.*, vol. 24, pp. 283–292, 1987.

[4] J. Kates, "A test suite for hearing aid evaluation," *J. Rehab. Res. Dev.*, vol. 27, pp. 255–278, 1990.

[5] J. Beerends, A. Hekstra, A. Rix, and M. Hollier, "(PESQ), the new ITU standard for objective measurement of perceived speech quality, part II - Perceptual model," *J. Audio Eng. Soc.*, vol. 50, pp. 765–778, 2002.

[6] A. Rix, M. Hollier, A. Hekstra, and J. Beerends, "PESQ, the new ITU standard for objective measurement of perceived speech quality, part I - Time alignment," *J. Audio Eng. Soc.*, vol. 50, pp. 755–764, 2002.

[7] R. Huber and B. Kollmeier, "PEMO-Q–a new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911, November 2006.

[8] O. Dyrlund, "Coherence measurements in hearing instruments, using different broadband signals," *Scand. Audiol.*, vol. 21, pp. 73–78, 1992.

[9] 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.*, vol. 122, no. 2, pp. 1150–1164, August 2007.

[10] C. Tan and B. Moore, "Perception of nonlinear distortion by hearing-impaired people," *International Journal of Audiology*, vol. 47, no. 5, pp. 246–256, 2008.

[11] W. Chu and Z. Ghahramani, "Preference learning with Gaussian processes," in *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

[12] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, January 2008.

[13] E. Jaynes, *Probability theory. The logic of science*. Cambridge, UK: Cambridge University Press, 2003.

[14] T. Dijkstra, A. Ypma, B. de Vries, and J. Leenen, "The learning hearing aid: common-sense reasoning in hearing aid circuits," *The Hearing Review*, vol. 14, no. 11, pp. 34–53, 2007.

[15] D. MacKay, "Bayesian methods for backpropagation networks," *Models of Neural Networks III*, pp. 211–254, 1994.

[16] C. Williams and D. Barber, "Bayesian classification with Gaussian processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1342–1351, 1998.

[17] D. MacKay, *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press, 2003.

[18] C. Rasmussen and C. Williams, *Gaussian processes for machine learning*. Cambridge, MA: MIT Press, 2006.

[19] C. Williams, "Prediction with Gaussian processes: from linear regression to linear prediction and beyond," in *Learning and Inference in Graphical Models*, M. Jordan, Ed. Kluwer, 1998.

[20] D. MacKay, "Introduction to Gaussian processes," in *Neural Networks and Machine Learning*, C. Bishop, Ed. Berlin: Springer-Verlag, 1998, pp. 133–165.

[21] B. Schölkopf and A. Smola, *Learning with kernels*. Cambridge, MA: MIT Press, 2002.

[22] F. Jäkel, B. Schölkopf, and F. Wichmann, "A tutorial on kernel methods for categorization," *Journal of Mathematical Psychology*, vol. 51, pp. 343–358, 2007.

[23] N. Pourmand, D. Suelzle, V. Parsa, Y. Hu, and P. Loizou, "On the use of Bayesian modeling for predicting noise reduction performance," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2009, pp. 3873–3876.

[24] C. Rasmussen, "Evaluation of Gaussian processes and other methods for non-linear regression," Ph.D. dissertation, University of Toronto, Canada, March 1996.

[25] I.-T. P.835, *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*, 2003.

[26] D. Kingsley, *Preference uncertainty, preference refinement and paired comparison choice experiments*. Dept. of Economics. University of Colorado, 2006.

[27] E. Brochu, N. de Freitas, and A. Ghosh, "Active preference learning with discrete choice data," in *Advances in Neural Information Processing Systems 20*, J. Platt, Y. Koller, D. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008, pp. 409–416.

[28] I.-T. R. P.800.1, *Mean opinion score (MOS) terminology*, ITU, Geneva, Switzerland, 2003.

[29] I.-T. R. P.830, *Subjective performance assessment of telphone-band and wideband digital codecs*, ITU, Geneva, Switzerland, 1996.

[30] W. Chu and Z. Ghahramani, "Gaussian processes for ordinal regression," *Journal of Machine Learning Research*, vol. 6, pp. 1019–1041, 2005.

[31] P. Groot, T. Heskes, and T. Dijkstra, "Nonlinear perception of hearing-impaired people using preference learning with Gaussian processes," Radboud University Nijmegen, Tech. Rep. ICIS-R08018, 2008.

[32] K. Train, *Discrete choice methods with simulation*. Cambridge University Press, 2003.

[33] R. Bradley and M. Terry, "Rank analysis of incomplete block designs, I. the method of paired comparisons," *Biometrika*, vol. 39, pp. 324–345, 1952.

[34] R. Luce, *Individual choice behaviours: a theoretical analysis*. Mineola, New York: Dover publications, 2005.

[35] 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.*, vol. 95, pp. 1085–1099, 1994.

[36] J. Kates and K. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, pp. 2224–2237, 2005.

[37] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, ser. Springer Series in Statistics. Springer, February 2009.

[38] B. Everitt, *The analysis of contingency tables*. London, UK: Chapman and Hall, 1977.

[39] T. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

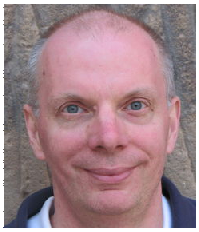
[40] S. Salzberg, "On comparing classifiers: pitfalls to avoid and a recommended approach," *Data Mining and Knowledge Discovery*, vol. 1, pp. 317–327, 1997.

[41] E. Bonilla, K. Ming, A. Chai, and C. Williams, "Multi-task gaussian process prediction," in *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., vol. 20. MIT Press, 2008.

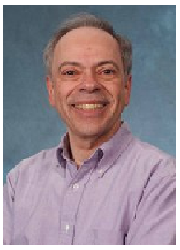
[42] A. Birlutiu, P. Groot, and T. Heskes, "Multi-task preference learning with an applications to hearing-aid personalization," *Neurocomputing*, 2009, doi:10.1016/j.neucom.2009.11.025.

Perry Groot received the M.Sc degrees in Computer Science (artificial intelligence) and Mathematics (topology) in 1998 and a Ph.D. degree from the Department of Artificial Intelligence, Vrije University Amsterdam in 2003. Currently, he is a researcher at the Institute for Computing and Information Sciences at the Radboud University Nijmegen, the Netherlands. His research interests include Bayesian reasoning applications, particularly the use of Gaussian processes for preference learning and function optimization.

Tom Heskes received both the M.Sc and Ph.D degrees in Physics from the University of Nijmegen, The Netherlands, in 1989 and 1993, respectively. After a year of postdoctoral work at the Beckman Institute in Urbana-Champaign, Illinois, he worked from 1994 to 2004 as a researcher for SNN, Nijmegen. Between 1997 and 2004 he headed the company SMART Research BV, specialized in applications of neural networks and related techniques. Since 2004 he works at the Institute for Computing and Information Sciences at the Faculty of Science, Radboud University Nijmegen, where he is now a full professor in artificial intelligence. Prof. Heskes is Editor-in-Chief of Neurocomputing. His research interests include theoretical and practical aspects of neural networks, machine learning, and probabilistic graphical models.



Tjeerd M.H. Dijkstra received his M.Sc degree in Physics from the University of Utrecht in 1988 and his Ph.D. in Biophysics from the University of Nijmegen in 1994. After a post-doc in Psychology at the University of Pennsylvania, he held a faculty position in Psychology at The Ohio State University. From 2005 until 2009 he worked at the hearing-aid company GN ReSound in Eindhoven. His research interests include human perception, both visual and auditory, human motor control, bioinformatics and systems biology.



James M. Kates was born in Brookline, Massachusetts, in 1948. He received the Bachelor of Science and Master of Science degrees in electrical engineering from the Massachusetts Institute of Technology in 1971, and the professional degree of Electrical Engineer from M.I.T. in 1972. He currently holds the position of Research Fellow with GN ReSound. He has an office in Boulder, Colorado, where he is involved in research in digital signal processing for hearing aids. He is also an adjunct professor in the Department of Speech Language and Hearing Sciences at University of Colorado in Boulder, where he teaches courses in hearing-aid technology and conducts research in auditory perception, hearing loss, and signal processing for hearing aids. Prior positions include Senior Research Scientist and adjunct faculty at the Center for Research in Speech and Hearing Sciences of the City University of New York, Director of Basic Research for Siemens Hearing Aids in Piscataway, NJ, and work in radio direction finding, speech coding and enhancement, digital audio, and loudspeaker design and evaluation. He is a Fellow of the Acoustical Society of America and a Fellow of the Audio Engineering Society, and is the author or co-author of 50 technical papers, seven book chapters, and holds 20 US patents. His book Digital Hearing Aids was published in 2008 by Plural Publishing, San Diego.