Nonlinear perception of hearing-impaired people using preference learning with Gaussian Processes

Perry Groot[1], Tom Heskes[1], Tjeerd M.H. Dijkstra[1,2], James M. Kates[4]

[1] Radboud University 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

[4] University of Colorado, Dept. Speech Language and Hearing Sciences. Boulder, CO

perry@cs.ru.nl, tomh@cs.ru.nl, t.dijkstra@cs.ru.nl, jkates@gnresound.dk

Abstract

A probabilistic kernel approach to pairwise preference learning based on Gaussian Processes is applied to speech 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. Arehart, et. al. (2007) [J. Accoust. Soc. Am. **122**(2), 1150–1164] reported pairwise comparisons for 14 normal-hearing and 18 hearing-impaired subjects for several sound distortions. Using the kernel approach gives a significant improvement in prediction results. We show a significant difference between normal-hearing and hearing-impaired subjects, because of nonlinearities in the perception of hearing-impaired subjects.

PACS numbers: 43.71.An, 43.71.Gv, 43.71.Ky, 43.66.Lj

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. 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] and coherence based methods.[5;6] Tan and Moore [7] have written 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 remain[5] and some model extensions have been surprisingly ineffective possibly because of random variability in the judgments of subjects.[7]

It is well-known, however, that hearing-impaired subjects appear to have only moderate test-retest reliability when judging sound quality[8;9] 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.[7] 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.[5] pairwise comparisons were performed with 14 normal-hearing and 18 hearing-impaired subjects for several sound distortions. Arehart et al.[5] 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[10]; (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 compares the classification results of Arehart et al. [5] with a linear and nonlinear classifier using GPs. Section 4 gives conclusions.

2 Bayesian Framework

Let $X = \{x_i, \ldots, x_n\}$ be a set of n distinct sound samples with $x_i \in \mathbb{R}^d$, i.e, sounds are represented by some of their 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 $v_{i,1}$ is preferred over $v_{i,2}$ and $d_i = -1$ otherwise. We define a prior over functions by assuming that 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 or kernel function. Here we use the Gaussian kernel (with parameter κ) and the linear kernel.[10] 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, includes a user response bias b that depends on the order of samples presented, and that the individual observations are contaminated with Gaussian noise, i.e., both $\delta_1, \delta_2 \sim \mathcal{N}(0, \sigma^2)$

$$p(v_{k,1}, v_{k,2}, 1|f(v_{k,1}), f(v_{k,2})) = p(f(v_{k,1}) + \delta_1 > f(v_{k,2}) + b + \delta_2) \tag{2}$$

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.

The posterior follows from Bayes' rule and is approximated using the Laplace approximation. Model selection is done by a maximum likelihood estimation of the hyperparameters $\{\kappa, \sigma, b\}$. For full details see Groot et al. [11] .

3 Empirical Results

We use data from Arehart et al. [5] , who collected pairwise preference data using listener experiments (14 normal-hearing and 18 hearing-impaired subjects). The stimuli presented were two sets (one male, one female talker) of concatenated sentences from the hearing-in-noise-test (HINT), [12] subjected to three types of degradation: symmetric peak-clipping, symmetric center-clipping, and additive stationary speech-shaped noise (each with 8 levels of degradation). Sound samples are represented using the coherence speech intelligibility index (CSII), giving three features CSII_{Low}, CSII_{Mid}, and $\text{CSII}_{\text{High}}$.

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.

For constructing classifiers, we use 10-fold cross-validation, which partitions a data set into 10

subsets. The classifier is then trained 10 times on 9 subsets and tested on the remaining subset. Each sample is used for training and used exactly once for testing. The results can be combined to produce a single estimate (e.g., the accuracy is the overall number of correct classifications, divided by the number of instances). In order to determine whether one classifier $\hat{f}_A$ significantly improves upon another classifier $\hat{f}_B$, McNemar's test can be used.[13] McNemar's test only focusses on the outcomes of the classifiers that are different. 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.

In this section we compare three classifiers on the pairwise comparison data of Arehart et al.[5]. The first classifier is the Q_3 metric reported by Arehart et al.[5], which is a minimum mean-squared error fit to the normal-hearing subjects' quality ratings given by

$$Q_3 = \frac{1}{1+e^{-c}} \text{ with } c = -4.56 + 2.41 \cdot \text{CSII}_{\text{Low}} + 2.16 \cdot \text{CSII}_{\text{Mid}} + 1.73 \cdot \text{CSII}_{\text{High}} \qquad (3)$$

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 we trained two other classifiers. A GP with a linear kernel (LK) and a GP with a Gaussian kernel (GK). The linear kernel corresponds to probit regression.[14] Both kernels were trained using 10-fold cross-validation and were compared to each other using the McNemar test. 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).(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).) 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 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 modelling 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.

To demonstrate the nonlinear perception in hearing-impaired subjects we show in Figure 1 the elicited utility function of one of the subjects with a significant improvement in prediction quality. We fitted the CSII_{Mid} features in terms of the CSII_{Low} and $\text{CSII}_{\text{High}}$ features as these were highly

correlated and to reduce our graph to 3-dimensions. Figure 1 shows the hyperplane in terms of CSII_{Low}, $\text{CSII}_{\text{High}}$, and regressed CSII_{Mid} features, the utility function on this hyperplane, and the samples projected on the contour plot of the utility function. Clearly, the elicitated utility functions shows nonlinear behaviour. Furthermore, we confirmed that the CSII approach significantly improved when incorporating the audiogram for computing sound features and that the nonlinear behaviour in perception is not just a byproduct of the CSII calculation approach.[11]

4 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, 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 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.

Acknowledgments

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.

References and links

[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.*, 49: 1011–1048, 2001.

[2] 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, 2001.

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

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

[5] K.H. Arehart, J.M. Kates, C.A. Anderson, and L.O. Harvey Jr. Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.*, 122(2):1150–1164, August 2007.

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

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

[8] A. Gabrielsen, B.N. 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.

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

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

[11] P.C. Groot, T. Heskes, and T.M.H. Dijkstra. Nonlinear perception of hearing-impaired people using preference learning with Gaussian Processes. Technical Report ICIS-R08018, Radboud University Nijmegen, 2008.

[12] M. Nilson, S.D. 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.

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

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

Figure captions

Table 1. Prediction error (PE) of several methods 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.

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).

Table 1:

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

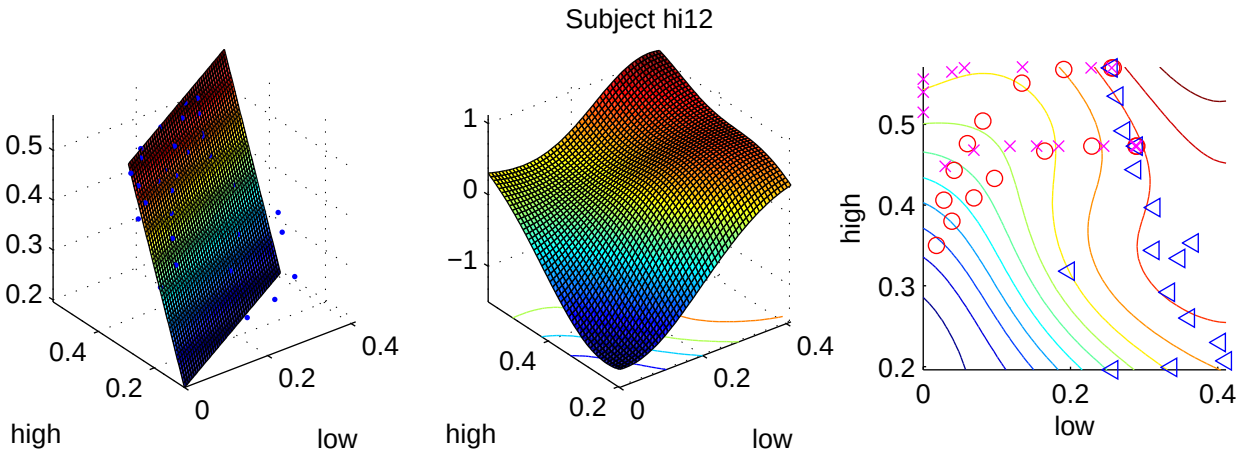

Figure 1: Color online!