Data Analysis Methods and its Applications

Edited by
Józef Pociecha & Reinhold Decker

Wydawnictwo C.H. Beck

Publisher: Dorota Ostrowska-Furmanek
Substantive editing: Dominika Drygas
German articles reviewer: Józef Pociecha
Polish articles reviewer: Reinhold Decker
Design of the cover and title pages: Maryna Wiśniewska
Image on the cover: ©MarkEvans/iStockphoto.com

Typeset by TEX

Wydawnictwo C.H. Beck Sp. z o.o.
ul. Bonifraterska 17, 00-203 Warszawa

Typesetting: Wydawnictwo C.H. Beck
Printed and bound by: Elpil Siedlce

ISBN 978-83-255-3458-5

Statistics for hearing aids: Auralization

Claus Weihs, Klaus Friedrichs, Bernd Bischl

Chair of Computational Statistics, TU Dortmund, Germany

e-mail: {weihs, friedrichs, bischl}@statistik.tu-dortmund.de

Abstract. Based on a computational auditory model, a method for quality measurement of sound transformation algorithms for hearing aids is studied by means of statistical inversion of the model. Auditory models describe the transformation from acoustic signals into spike firing rates of the auditory nerves by emulating the signal transductions of the human auditory periphery. The inverse approach, which is called auralization, is discussed in this paper.

There have already been few successful attempts to auditory inversion each of which deal with relatively simple auditory models. In recent years more comprehensive auditory models have been developed which simulate nonlinear effects in the human auditory periphery. Since for this kind of models an analytical inversion is not possible, a statistical auralization approach using classification and regression methods is proposed.

Keywords: auditory model inversion, classification, regression, MARS.

1. Introduction

Measuring the quality of signal algorithms for hearing aids is a hard challenge. Tests with probands are difficult and expensive because there are many different kinds of hearing impairments. Cheaper and simpler are tests using an auditoy model, i.e. a computer model of the human auditory system. It requires an acoustic signal as input and outputs the spike firing rates of the auditory nerve fibers. The human auditory system consists of roughly 3000 auditory nerve fibers but in auditory models this is usually simplified to a much smaller quantity. A popular model is the one of Meddis and Sumner [8], which has also been used for this study. In this model the auditory system is coded by a multichannel bandpass filter where each channel represents one specific nerve fiber. As in the human system each channel has its specific center frequency by which the perceptible frequency range is defined. Figure 1 shows the center frequencies of the 30 channels, as they are defined in the default settings of the model. The output of the model, called auditory image, can be seen in figure 2. While the 30 channels are located on the

vertical axis and the time response on the horizontal axis, the greyscale indicates the spiking probability per second.

After having implemented some kind of hearing loss in the model, as introduced by Jepsen [4], auditory images of the modified model can be compared to the ones of the normal-hearing model. While a hearing aid would be perfect if it would produce exactly the output of the normal-hearing model for the hearing-impaired listener as well, this requirement is almost impossible for most kinds of hearing impairments. Thus, the distance between two auditory images has to be measured. Unfortunately, it is not known how the human brain interprets the auditory images and so defining a distance function is very complicated. Instead, in this paper the inverse procedure to resynthesize the original signal from the auditory image is proposed. Due to this method the auditory image of the hearing-impaired listener gets hearable and can be compared easily by a sound test to the original signal. More precisely, it is even not essential to get the input signal exactly but it is sufficient to get a signal which sounds like it. This procedure is called auralization.

There have already been successful attempts of auralization by analytical inversion. Slaney introduced techniques to recreate sounds from perceptual displays known as cochleagrams and correlograms [7] and Hohmann presented an approach to invert the gammatone filter bank, a filter which is also used in Meddis auditory model [3]. Feldbauer analyzed the problem from another direction. He developed an auditory model with the intention that it can be inverted with a relatively low computational effort [1].

In the more comprehensive model of Meddis cochlear-nonlinearities are modeled which are important with respect to many perceptual experiments and animal observations. Unfortunately, the resulting model is impossible to invert analytically. Therefore, in this study an auralization approach using statistical methods is introduced.

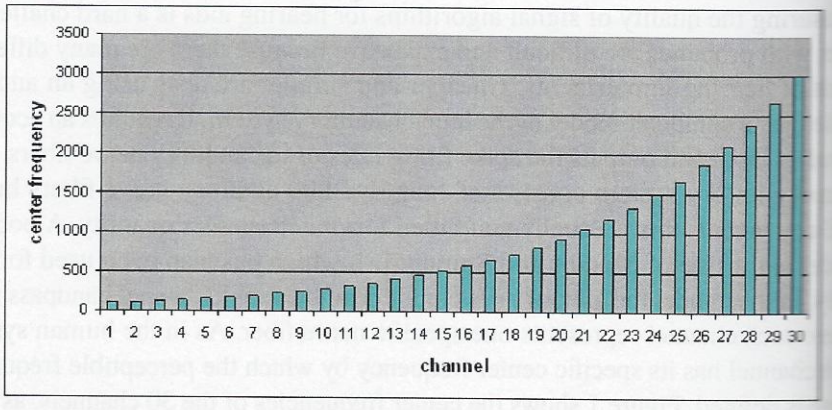


Figure 1. Center Frequencies (CF) of Meddis model

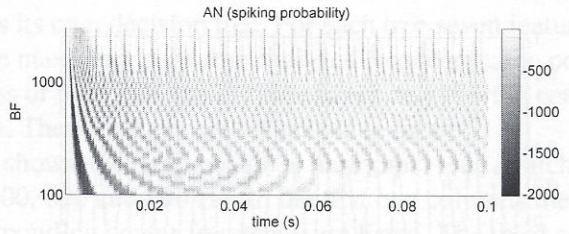


Figure 2. Auditory image

2. Auralization by using statistical methods

To simplify the problem in this study it is assumed that the input stimulus is one harmonic tone. Harmonic tones are typical signals for musical sound. They consist of a fundamental frequency, i.e. the key tone, and integer multiples of this frequency, which are called overtones. The key tone and the overtones together are called partial tones. The sound of a harmonic tone is defined by its involved frequencies and the power of each frequency. Figure 3 shows an exemplary harmonic tone which contains a key tone of 300 Hz and the overtones of 600 Hz and 900 Hz. Additional to the contained frequencies the sound of a tone is also dependent on their power. In the example the key tone (300 Hz) has a power of 80 dB, the first overtone (600 Hz) a power of 87 dB and the second overtone (900 Hz) a power of 77 dB.

Thus, to resynthesize a harmonic tone a two-stage concept is proposed. In a first step the key tone and all involved overtones have to be detected by classification and in a second step the power of each partial tone has to be estimated by regression.

A crucial method for the whole task is to use the phase locking effect which was introduced by Moissl and Meyer-Base [5]. This effect phase-locks the impulse rate of the channels to the stimulus. In our problem this implies that each frequency, which is part of the input signal, will also occur in some channels. This correlation can be seen in figure 4, where an input stimulus and the corresponding auditory image are plotted. Exemplarily, a frequency of 300 Hz, which means 3 periods every 0.01 seconds, can be discovered in both images. In the auditory image this period can be seen at the top most clearly.

2.1. Frequency detection

In frequency detection the periodicities of each channel output have to be analyzed. Therefore, the discrete Fourier transform (DFT) of each channel output is generated. Because of the phase locking effect in each of these outputs peaks should occur at frequencies which are part of the acoustic stimulus. Such a peak gets the stronger the smaller the difference is between this frequency and the center frequency of the channel. From this it follows that it is sufficient to detect in each channel only the neighboring frequencies of its center frequency. Furthermore,

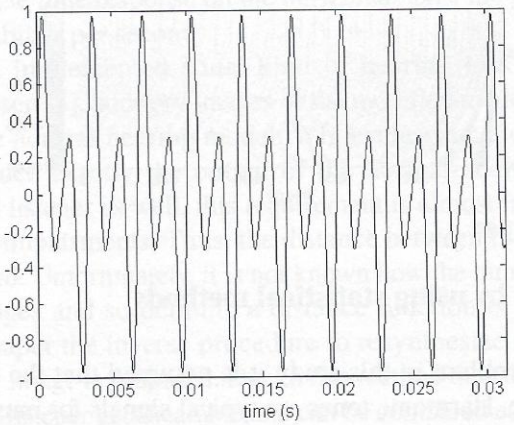


Figure 3. Exemplary harmonic tone: key tone: 300 Hz (80 dB), 1st overtone: 600 Hz (87 dB), 2nd overtone: 900 Hz (77 dB)

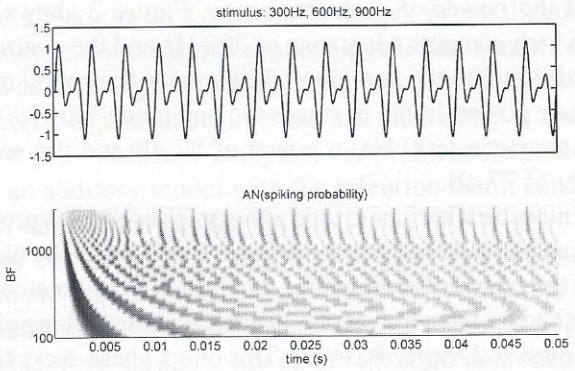


Figure 4. Phase locking effect: Frequencies in the input occur in the auditory image as well

the restriction on harmonic tones ensures that in each of the 30 channels at most one frequency has to be detected.

Figure 5 shows the DFT of channel 13 for a harmonic tone which includes an overtone with the frequency of 400 Hz. This frequency is also the maximum peak of this chart. A first approach is detecting the main peaks of all channels and, in this way, getting all frequency components. Unfortunately, figure 6 shows that the maximum peak does not always define a frequency which is part of the acoustic stimulus. Here the maximum peak is also at 400, but in this example 400 Hz is not part of the input tone. Therefore, after having detected the maximum peak of the DFT it has to be classified if this frequency is in fact part of the original signal in order to construct a classification rule. In this study this is done by classification trees. Therefore, to enable supervised learning a training set of harmonic tones is required. Because there are differences between low and high frequencies each

channel needs its own decision tree. For each tree seven features are used: The location of the maximum peak (the analyzed frequency), the power of this peak, the smoothness of this peak, the distance to the neighboring peaks and the power of these peaks. These features are visualized in figure 7.

Figure 8 shows the features of a harmonic tone which consists of the frequencies 300, 600 and 900 Hz. In the first two columns the channel number and its corresponding center frequency are listed. The third column shows the detected main frequency of each channel. The features, which are used for the classification task, are listed in columns 4–9. Finally, the last two columns show the target variables for the classification respectively the regression task which enable supervised learning. An exemplary classification tree is shown in figure 9. In this example the probability for "Yes" i.e. that the frequency is a tone component is higher if the maximum peak is high, the location of this peak is low, the peak of the next frequency on the right side is low or the distance to the next peak on the left side is low.

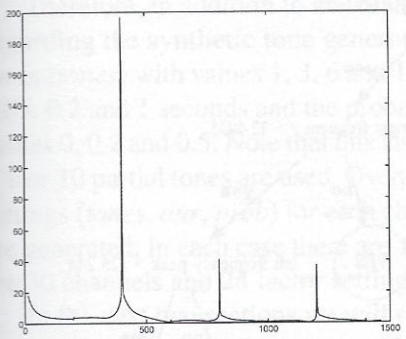


Figure 5. DFT of channel 13: 400 Hz is part of the tone

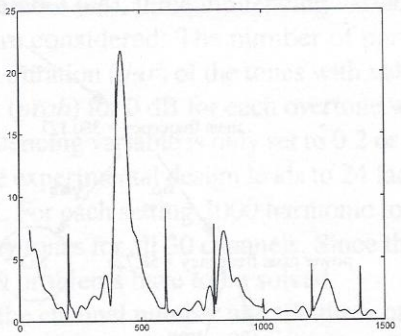
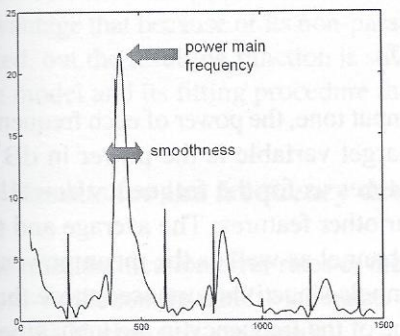


Figure 6. DFT of channel 13: 400 Hz is not part of the tone



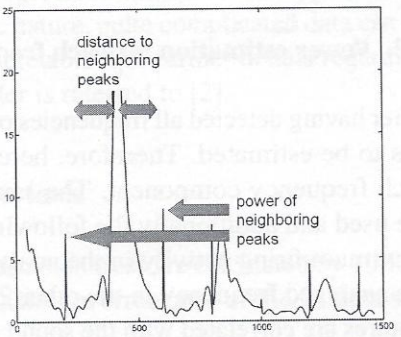


Figure 7. Features

channel	center frequency	main frequency	power main frequency	smoothness	power left	distance left frequency	power right frequency	distance right frequency	frequency is tone component?	power of frequency in dB
11	296	301.8	185.9	23.4	17.9	285,6	32,4	599,1	**yes**	82
12	333	312.0	18.7	474.6	17.7	295,9	32,9	588,9	**no**	
13	375	360.4	7.9	606.4	17.2	344,2	51,6	540,5	**no**	
14	422	449.7	6.8	741.2	16.2	433,6	52,7	751,5	**no**	
15	475	505.4	7.7	609.4	16.6	489,3	50,6	695,8	**no**	
16	535	568.4	10.6	720.7	16.5	552,2	54,7	632,8	**no**	
17	602	602.1	151.6	39.6	16.4	585,9	49,7	599,1	**yes**	87
18	677	701.7	12.0	796.9	14.6	685,5	122,5	199,2	**no**	
19	762	810.1	5.0	1063.5	14.8	793,9	51,8	991,7	**no**	
20	857	900.9	131.1	41.0	15.4	298,8	53,5	900,9	**yes**	77

Figure 8. Feature generation of an exemplary harmonic tone

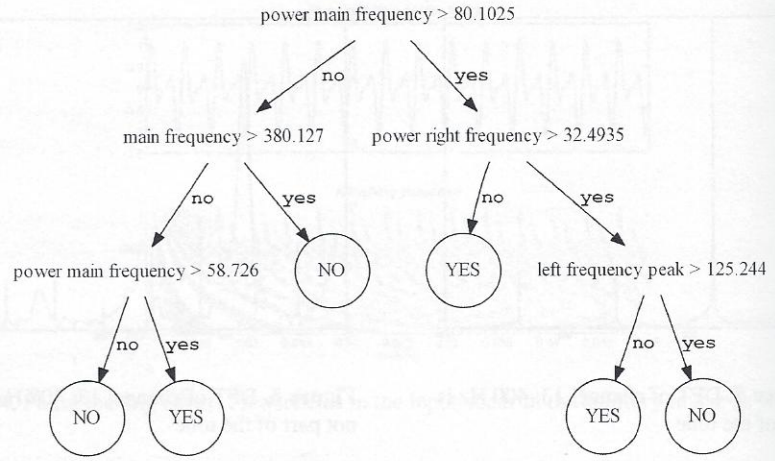


Figure 9. Classification tree of channel 13

2.2. Power estimation for each frequency

After having detected all frequencies of the input tone, the power of each frequency has to be estimated. Therefore, here the target variable is the power in dB of each frequency component. The same features as for the frequency detection are used and additionally the following four other features: The average and the maximum firing activity in the analyzed channel as well as the mean power of the analyzed frequency in the other 29 channels since it is supposed these three features are correlated with the sound volume of the frequency in the input signal. Furthermore, the overtone number of the frequency in the harmonic tone is used which is known after all involved frequencies are detected. Since this problem

should be more complicated than the one in section 2.1, 7 regression methods are compared: a simple linear model, a linear model with two-way interactions, a complete second-order linear model, a regression tree, a random regression forest, a kriging model with a Matern 5/2 covariance kernel and a kriging model with a Gaussian covariance kernel. For kriging see [6], for the other models [2].

3. Experimental design

To test the approach for different kinds of harmonic tones an experimental design generating several training sets of synthetic tones is used. Generally for all tones the key tone has a frequency between 80 and 3400 Hz. The powers of the key tone and the first overtone are uniformly distributed between 60 and 95 dB, whereas the power of each other overtone is distributed between 55 and 85 dB according to a beta distribution. With a specified probability the power of these overtones can also be 0 dB.

Therefore, in addition to the channel number (*ch*), three influencing variables regarding the synthetic tone generation are considered: The number of partial tones (*tones*) with values 1, 3, 6 and 10, the duration (*dur*) of the tones with values 0.05, 0.2 and 1 seconds and the probability (*prob*) for 0 dB for each overtone with values 0, 0.2 and 0.5. Note that this last influencing variable is only set to 0.2 or 0.5 if 6 or 10 partial tones are used. Overall, the experimental design leads to 24 factor settings (*tones*, *dur*, *prob*) for each channel. For each setting 3000 harmonic tones are generated, in each case there are 100 key tones for all 30 channels. Since there are 30 channels and 24 factor settings, 720 problems have to be solved.

In the next two sections we will relate the channel number and the three other influencing factors to the misclassification error rate of the decision tree and the root mean square error of the regression models, respectively. For this, we fit multiple adaptive regression splines (MARS) to our experimental data in order to perform a further analysis regarding which factors and interactions have the most effect on the performance value. A regression model of this type has the advantage that because of its non-parametric nature, quite complicated data can be fitted, but the resulting function is still interpretable. For further details regarding the model and its fitting procedure the reader is referred to [2].

4. Results for the frequency detection task

The misclassification error rates of the 720 decision trees are calculated by 10-fold cross-validation. In table 1 the mean and maximum error rates of the 24 data sets over all 30 channels are listed.

As can be seen in the upper two plots of figure 10 of the MARS model, the error rate is mainly dependent on the number of partial tones and the duration of

the input tones. The first effect is expected, since the more frequencies a signal contains, the more complex (and harder to classify) it gets. Shorter tones perform worse because a higher error occurs in calculating the discrete Fourier transform for low frequencies. For example in 0.05 seconds the frequency 100 Hz just has 5 periods. In contrast, the probability for 0 dB for each overtone does not seem to have much influence on the result. Maybe this is because a higher *prob* value results in more diverse tones, but also on average in a lower number of partial tones. Taken together, these two aspects might compensate. Table 2 deals with the mean error rates of the 30 channels over all 24 data sets. As it can be seen here the error rates are higher for the upper channels on average. Figure 10 reveals that this is specifically true, if the number of partial tones is also high.

This fact is probably due to the employed method of tone generation. In the lower channels a frequency can only occur as a key tone which ensures a much lower complexity than in the upper channels, in which each frequency can be the key tone as well as any overtone. In fact this impact is insignificant for signals without overtones, as each contained frequency can just be the key tone even in the upper channels. Note that the shown MARS model has an R^2 of 0.80.

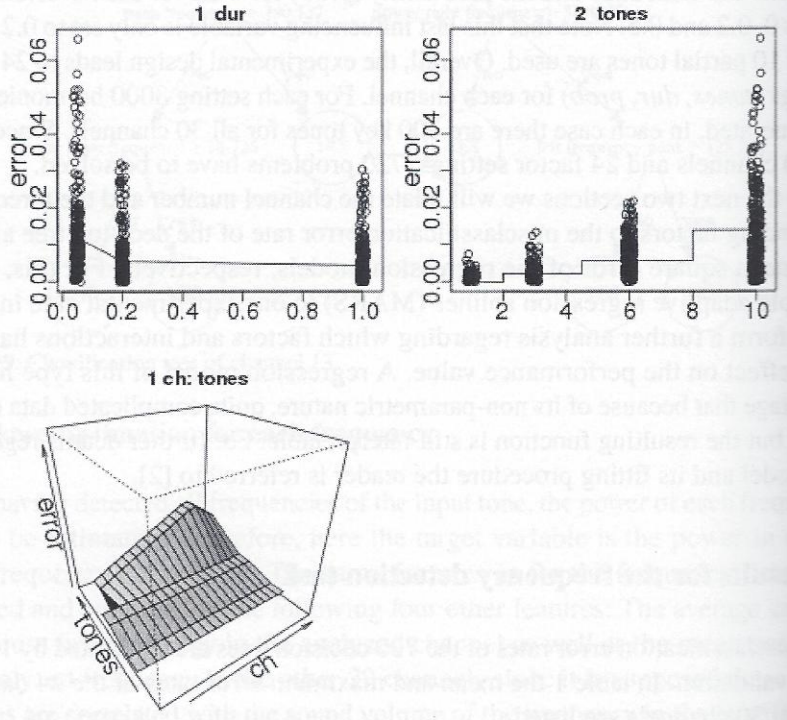


Figure 10. Effects plot for MARS model, relating *ch*, *tones*, *dur* and *prob* and to the misclassification error of the decision tree. Note that only the most important effects are shown

Table 1. Mean and maximum error rates of frequency detection over all 30 channels

	Mean error rate in %	Max error rate in %
1 partial tone, 1 sec	0.1	0.3
1 partial tone, 0.2 sec	0.1	0.5
1 partial tone, 0.05 sec	0.2	0.7
3 partial tones, 1 sec	0.2	0.7
3 partial tones, 0.2 sec	0.2	0.6
3 partial tones, 0.05 sec	0.6	1.3
6 partial tones, 1 sec, 0prob = 0	0.4	0.8
6 partial tones, 0.2 sec, 0prob = 0	0.5	1.0
6 partial tones, 0.05 sec, 0prob = 0	1.3	2.2
6 partial tones, 1 sec, 0prob = 0.2	0.6	1.2
6 partial tones, 0.2 sec, 0prob = 0.2	0.6	1.1
6 partial tones, 0.05 sec, 0prob = 0.2	1.1	1.9
6 partial tones, 1 sec, 0prob = 0.5	0.5	1.1
6 partial tones, 0.2 sec, 0prob = 0.5	0.5	1.2
6 partial tones, 0.05 sec, 0prob = 0.5	1.0	1.8
10 partial tones, 1 sec, 0prob = 0	0.6	1.4
10 partial tones, 0.2 sec, 0prob = 0	1.2	2.8
10 partial tones, 0.05 sec, 0prob = 0	2.6	6.6
10 partial tones, 1 sec, 0prob = 0.2	1.2	3.0
10 partial tones, 0.2 sec, 0prob = 0.2	1.4	3.2
10 partial tones, 0.05 sec, 0prob = 0.2	2.3	5.0
10 partial tones, 1 sec, 0prob = 0.5	0.9	2.4
10 partial tones, 0.2 sec, 0prob = 0.5	0.9	2.1
10 partial tones, 0.5 sec, 0prob = 0.5	1.4	2.7

Table 2. Mean error rates of frequency detection over all 24 data sets

Channel number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Error rate in %	0.2	0.4	0.5	0.5	0.3	0.3	0.3	0.4	0.6	0.7	0.6	0.5	0.7	0.7	0.6
Channel number	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Error rate in %	0.9	1.0	1.2	1.4	1.3	1.4	1.3	1.4	1.4	1.3	1.4	1.3	1.1	1.0	1.1

5. Results for the power estimation task

The seven regression methods are applied to the 720 problems and the root mean square error (RMSE) is calculated by using 10-fold cross-validation. The minimum, median and maximum error rates of each method – aggregated over all problems – are listed in table 3. These results should be related to the power range of each frequency which lies between 55 dB and 95 dB as mentioned in section 3.

This means that the standard deviation of each problem is around 9 dB. It can be seen that the RMSE ranges from fine 0.04 dB to worse 3.7 dB. However, as will be shown in the next section, even the worst obtained deviation in our experiments is not audible for the human ear in many harmonic tones. A comparison of the seven regression methods reveals that the kriging model with Matern 5/2 covariance kernel has the best median error, the kriging model with the Gaussian kernel and the quadratic linear model have the lowest minimum error and random forest has the best maximum error.

Now each regression model is assigned a ranking score for each of the 720 problems and the result is summarized in figure 11. Kriging with the Matern kernel is the best method for most of the problems. But also the random forest and kriging

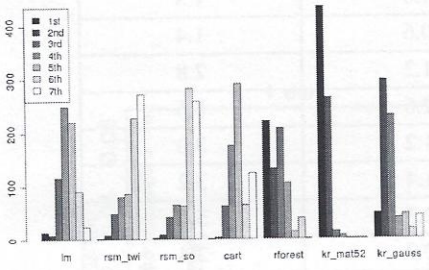


Figure 11. Ranking of the 7 regression methods summarized over all 720 problems

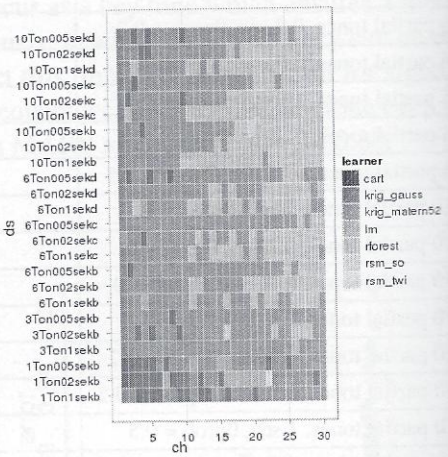


Figure 12. Best model for each problem

with a Gaussian kernel perform often quite well. In contrast, the other 4 models are only rarely best. Figure 12 shows which regression model is best for which problems. The channels are enumerated on the horizontal and the data sets on the vertical axis. As described in section 3 the data sets are defined by their number of partial tones (1, 3, 6 and 10), their tone duration (0.05, 0.2 and 1 sec.) and their 0 dB probability for each overtone (b=0, c=0.2 and d=0.5). While for lower channels and only one partial tone kriging with the Gaussian kernel is best, for upper channels and more complex tones (6 and 10 partials) random forest has the lowest error. For most of the other problems kriging with the Matern kernel performs best. Figure 13 shows the RMSE for each problem when selecting the best model. As expected for few overtones a lower RMSE can be obtained, which can also be seen in figure 14, where the average RMSE of data sets containing the same number of partial tones are plotted versus the channel number. Furthermore, it can be noticed that the

lower channels perform better, an effect which was already observed in the results regarding the frequency detection. In contrast to frequency detection, here the error surprisingly gets the smaller the shorter the tones are, a fact which is confirmed by figure 15, which shows the RMSE densities for the 3 different duration values. Tones which have a duration of 0.05 seconds score best and 1 second worst. Thus, the important information about the power of a tone might be at the beginning of the spike activities. This observation could lead to future improvements of the proposed algorithm by using additional features which describe the beginning of a signal. In figure 16 the RMSE densities for the 3 different values of *prob* are

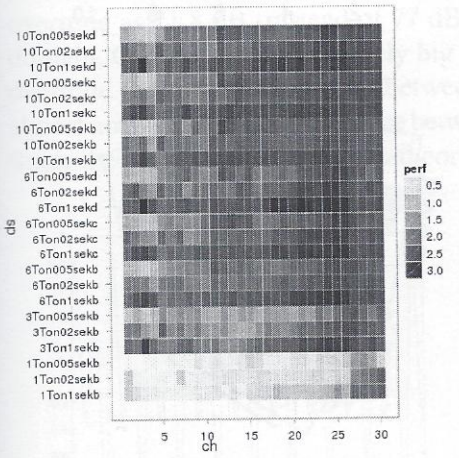


Figure 13. Best Performance for each problem

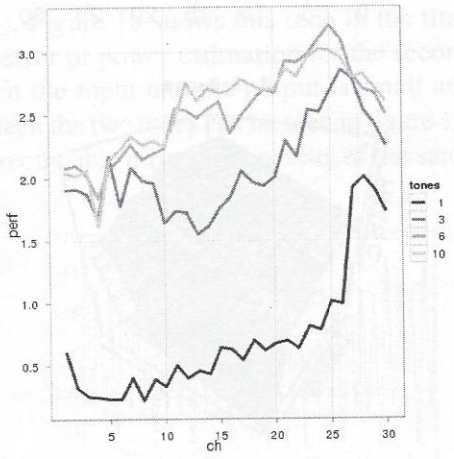


Figure 14. Performance dependent on the number partial tones

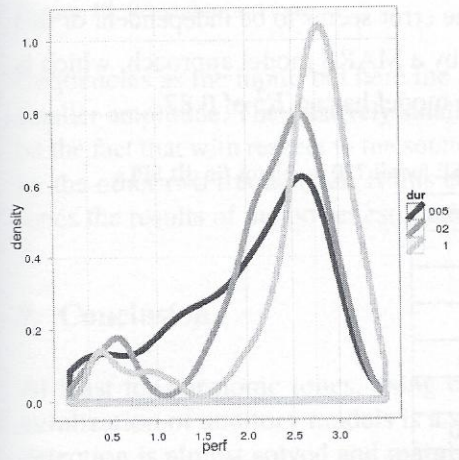


Figure 15. Performance dependent on tone duration

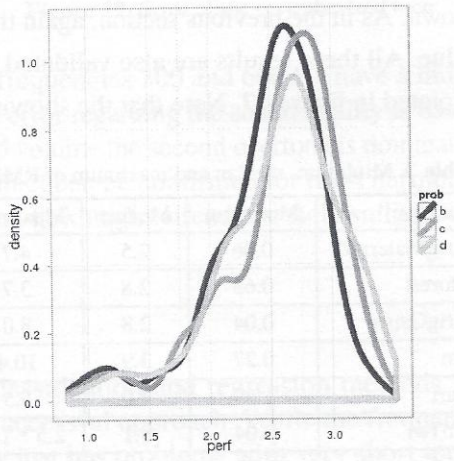


Figure 16. Performance dependent on 0 dB probability b = 0, c = 0.2, d = 0.5

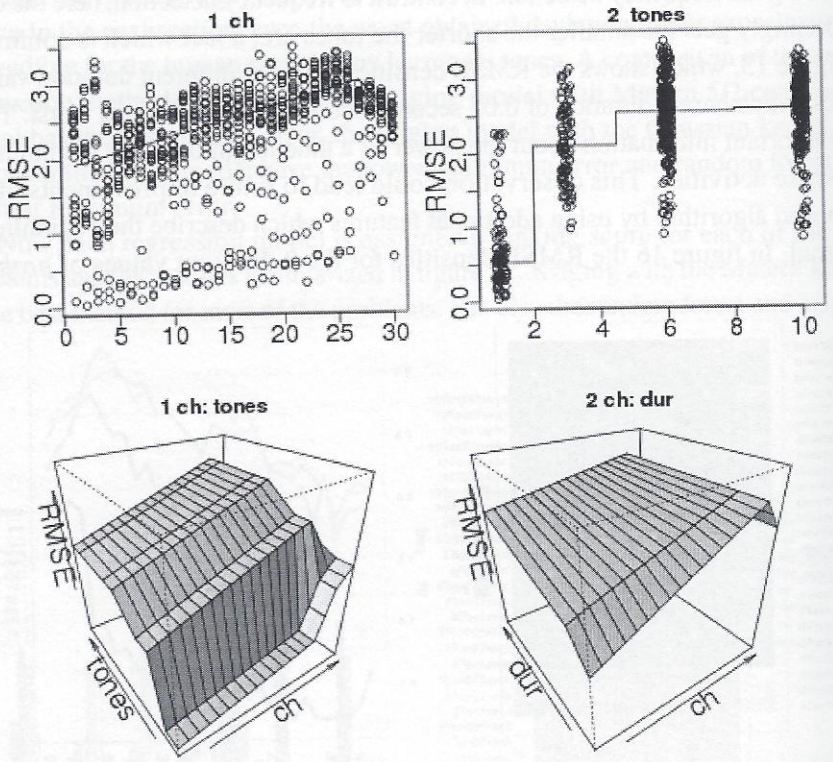


Figure 17. Effects plot for MARS model, relating *ch*, *tones* and *dur* and to the RMSE of the best regression model

shown. As in the previous section, again the error seems to be independent of this value. All these results are also validated by a MARS model approach, which is depicted in figure 17. Note that the shown model has an R^2 of 0.87.

Table 3. Minimum, median and maximum of RMSE for all 720 problems (in dB SPL)

Learner	Minimum	Median	Maximum
krigMatern52	0.06	2.5	4.7
rforest	0.63	2.8	3.7
krigGauss	0.04	2.8	8.0
lm	0.27	3.9	10.4
cart	1.57	4.0	5.5
lmTwi	0.04	8.1	$2.5 * 10^{11}$
lmSo	0.07	9.3	$3.6 * 10^{5}$

6. Resynthesizing test

While each partial tone and its sound volume can be detected relatively exact as it is shown in the last two sections, there does not exist a criterion for measuring the overall quality of auralization. Instead in this section an exemplary resynthesizing test of one harmonic tone is shown. As input the tone is used, which was already shown in section 2 respectively figure 3. The output tone after auralization consists of a key tone (301 Hz) and two overtones (602 Hz and 903 Hz). Since a difference between 300 Hz and 301 Hz is not hearable for humans, all frequencies are detected almost perfectly. The power of the key tone is estimated as 79.4 dB (instead of 80 dB), the first overtone as 86.7 dB (instead of 87 dB) and the second overtone as 71.8 dB (instead of 77 dB). Figure 18 shows this tone in the time domain. Contrary to the relatively big error of power estimation for the second overtone the overall difference between the input and the output is small and almost not hearable. The difference between the two tones can be seen in figure 19. Since the partial tones are detected correctly, this error plot consists of the same

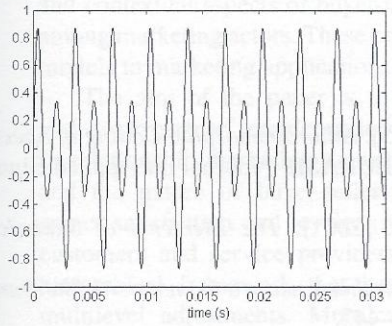


Figure 18. Resynthesized tone

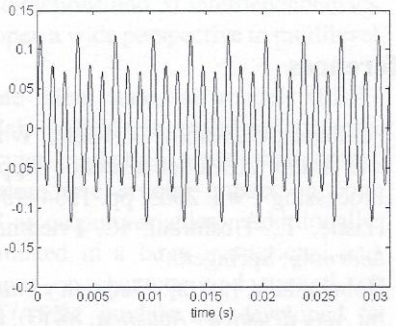


Figure 19. Error of the resynthesized tone

frequencies as the input, but here the frequencies 300 and 600 Hz have a much smaller amplitude. The relatively small error regarding the sound quality is based on the fact that with respect to the sound volume the second overtone is dominated by the other two frequencies. If this effect can be confirmed for most harmonic tones the results of the power estimation task might already be fully sufficient.

7. Conclusion

At least for harmonic tones, using classification and regression methods for auralization of auditory models is a successful approach. While the frequency detection is almost solved and mainly just has problems with very short tones the power estimation of each frequency component probably also leads to acceptable results. Since the results of this task are better for shorter tones a further

improvement could result from adding new features which just describe the beginning of a tone.

Current studies try to show the ability to generalize the proposed auralization approach to real tones and to the impaired-listener model. Auralization of this model can show which information is lost if some specific hair cells are damaged. Furthermore, a criterion has to be found to measure the overall quality of auralization. In this study this was done for each partial tone separately while the quality of the resynthesized tone had to be tested by sound tests. Finally, in future studies the auralization approach should also be adapted to more comprehensive input signals, in which power and pitch changes occur.

Acknowledgement

This work was supported by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823) and by the Research Training Group 'Statistical Modelling' of the German Research Foundation (DFG).

References

[1] Feldbauer, C., Kubin, G., Kleijn, W.B. [2005]: *Anthropomorphic Coding of Speech and Audio*: A Model Inversion Approach. EURASIP "Journal on Applied Signal Processing", vol. 2005, pp. 1334–1349.

[2] Hastie, T., Tibshirani, R., Friedman, J.H. [2001]: *The Elements of Statistical Learning*, Springer.

[3] Hohmann, V. [2002]: *Frequency analysis, synthesis using a Gammatone filterbank*. In: *Acta acustica / Acustica*, 88 (3), pp. 433–442.

[4] Jepsen, M.L., Dau, T., Ewert, S. [2006]: *A model of the normal, impaired auditory system*. Academic dissertation, Technical University of Denmark.

[5] Moissl, U., Meyer-Base, U., *Decoding of neural firing to improve cochlear implants*. In: Proc. SPIE, vol. 4055, pp. 337–349.

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

[7] Slaney, M., Naar, D., Lyon, R.F. [1994]: *Auditory model inversion for sound separation*. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, Adelaide, Australia, pp. 77–80.

[8] Sumner, C.J., O'Mard, L.P., Lopez-Poveda, E.A., Meddis, R. [2002]: *A revised model of the inner-hair cell and auditory nerve complex*. "Journal of the Acoustical Society of America", pp. 2178–2189.